TM12-260

DEPARTMENT OF THE ARMY TECHNICAL MANUAL

ARMY PERSONNEL TESTS AND MEASUREMENT

DEPARTMENT OF THE ARMY • APRIL 1953

AGO 2930A—Mar

DEPARTMENT OF THE ARMY TECHNICAL MANUAL

TM 12–260

This manual supersedes TM 12–260, 26 April 1946

ARMY PERSONNEL TESTS AND MEASUREMENT

DEPARTMENT OF THE ARMY • *APRIL 1953*

United States Government Printing Office
Washington: 1953

AGO 2930A

DEPARTMENT OF THE ARMY

WASHINGTON 25, D.C., *9 April* 1953

TM 12–260 is published for the information and guidance of all concerned.

[AG 201.6 (10 Mar 52)]

BY ORDER OF THE SECRETARY OF THE ARMY:

J. LAWTON COLLINS
Chief of Staff, United States Army

OFFICIAL:
WM. E. BERGIN
Major General, USA
The Adjutant General

DISTRIBUTION:

Active Army:
Tech Svc (1); AFF (3); OS Maj Comd (3); Base Comd (2); MDW (2); A (5); Tng Div (3); FT (3); Sch (5); GH (2); RTC (3); Pers Cen (3); RC (3); Rctg Main Sts (3).

NG: None.

Army Reserve: None.

For explanation of distribution formula, see SR 310–90–1.

CONTENTS

FOREWORD

The purpose of this manual is to provide an understanding of how the Army applies personnel psychology and statistics to its personnel problems.

This manual is addressed to two audiences. The first is composed of those officer and enlisted personnel whose assignments entail responsibility for using personnel measurement techniques and procedures or for instructing others in their use, either in schools or on the job. The second audience consists of officer and enlisted personnel interested in gaining a better understanding of technical aspects of personnel management.

Technical content of the manual is treated in the light of practical military personnel problems. The solutions to these problems are difficult; therefore, considerable study and thought will be required of the reader. So far as possible, technical terms have been avoided, but where such terms are used, their meanings are explained.

The manual emphasizes basic principles involved in Army personnel problems. Theoretical considerations have been included only when necessary to an understanding of Army personnel problems and the techniques employed in dealing with them. This is a manual on Army personnel psychology, not on general personnel psychology. Specific instructions for the various personnel measurement instruments used by the Army are contained in appropriate special regulations.

Throughout the text, where the word "man" is used, it may be understood to mean "man or woman." The same principles govern testing of both men and women, and in many classification and selection programs the same or similar tests are used. Different tests occasionally are used when the different childhood interests or similar considerations make such actions desirable.

This manual supersedes TM 12–260, 12 April 1946

CHAPTER 1

PERSONNEL MEASUREMENT AS AN AID IN PERSONNEL MANAGEMENT

Section I. THE EFFECTIVE UTILIZATION OF MANPOWER

1. General

An effective fighting force needs the right kind of man as well as the right kind of equipment. The right kind of man is the man who is suited to his job; he meets the requirements and he is not wasted. If it is known what the requirements are and what the characteristics of the men are, the jobs and men can be matched.

a. Effective Utilization of Manpower Means Matching Jobs and Men. It does not mean that only the best possible men will be accepted. The manpower barrel has a bottom and the "cream of the crop" is at best only a thin layer. How far down the barrel it is necessary to go is a matter of high-level policy and depends on how great the need is. It also depends on how successfully the men who are taken are used. The Army needs to know how the manpower barrel is made up. It is the over-all objective of personnel measurement to provide the essential information on the abilities which make up the Army's manpower.

b. Army Personnel Research Is Concerned With Discovering Techniques That Will Help Match Army People With Army Jobs. On the one hand, each Army job must be analyzed into the tasks which make it up and the skills needed to do the job. On the other, the abilities and skills that men and women bring with them from civilian life, or that they acquire in the Army, must be identified and described.

2. Analyzing Job Requirements

New occupations are constantly being developed to meet the changing needs of the Army. There are over 500 types of occupation in the Army, all of which have been established out of practical necessity. Once a job is recognized as essential in the work of a unit, it is added to the table of organization and a job description is worked out for it. The job description defines what a man is expected to do on the job and the degree of proficiency required. Selection standards for the job are established, and training courses organized, when required, to enable Army personnel to meet the job requirements.

3. Differences Among Army Personnel

a. General. The men and women who must be selected and assigned to Army training courses and jobs vary in many ways. The fact that each individual possesses more of some abilities than of others is important to the Army. Effective utilization of manpower is not possible without knowledge of the strengths and weaknesses of the individuals that make up the manpower pool. It is a severe loss to the Army to fail to use to advantage a man capable of developing into a good leader, just as it is a severe loss to place a man in a position of leadership and find out too late that he is incapable of carrying out his responsibilities.

b. Differences in Physical Characteristics. Even after screening by the physical examination at the induction station and subsequent physical conditioning, soldiers differ widely in health, strength, size, and endurance. Some men can march 30 miles a day with full field equipment; others can cover only a few miles under the same conditions. Some can resist extremes of temperature, others cannot; some can maintain their efficiency at high altitudes, others lose it; some can see well at night, others

are practically night blind. No matter how effective the screening or conditioning, soldiers will not show any great uniformity in physical characteristics.

c. Differences in Psychological Characteristics—Ability and Personality. The differences among soldiers in abilities and personality characteristics are just as large and as important as are the differences in physique, stamina, and the keenness of their senses. However, these abilities and personality characteristics cannot be observed as directly as can physical traits. It is not possible to tell by looking at a man whether he can add or spell, repair a carbine, lead a squad, or be aggressive under fire. Nor is it possible to tell by such direct observation whether the man can learn to do these things in the relatively short time available for training. To obtain useful estimates of psychological characteristics, it is necessary to use other methods—the methods which are described in this manual.

d. Differences in Effects of Training. Among Army men in the same training situation or in the same job the range in ability, and in capacity to absorb training, may be tremendous. Figure 1 illustrates the spread of men in one training unit rated according to the quality of their performance in the unit. Although all of these men had been subjected to the same training methods and schedule, they were not all rated as of equal value. Some were rated as of little value, some were rated as of "average" value, and some were rated as outstanding. They differed in their capacity to acquire the skills in which they were being trained. Men differ also in the level of skill they can reach with training. For instance, it is exceedingly doubtful if the poorest performers could be brought up to the level of the best even if limitless time were available. Differences among men are not always ironed out by training. Differences in level of performance may even increase under a program of instruction. Those who are more skillful to start with may improve faster than the less skillful, with the result that the range of abilities after training may be even greater than before.

4. Classification a Continuing Process

The jobs which the Army must fill rarely, if ever, require exactly the same abilities and aptitudes as are possessed by the men available for assignment. Since the Army requirements must be met, it is necessary that they take precedence over the soldier's abilities and interests whenever there is a conflict. The Army frequently will have to use men in duty assignments in which their abilities are not fully utilized. Later, as requirements change with the tactical situation, there may be a demand for more men with their level of ability, and consequently, they may be shifted to more appropriate jobs. Effective utilization of manpower means that such men are not lost sight of, that their superior abilities are known, recorded, and whenever possible, utilized.

5. The Army's Needs and Civilian Specialties

a. Many persons coming into the Army would like to perform jobs corresponding to those they performed in civilian life. Very often, since a man's work experience is carefully considered in the classification process, that is exactly what happens. Sometimes, however, there is a conflict between the needs of the Army and a man's expectation that he will be assigned to a particular occupation in which he believes—and with reason—that he will do best. In that case, the individual's desires must give way to the over-all needs of the Army. It must be recognized that the Army jobs to be filled and the men qualified to fill them never match exactly. Some compromise must always be made, and this compromise, of necessity, will be in the direction of adjusting the supply of manpower to the requirements of the Army. A typical assignment problem will show how such conflicts usually are resolved—To fill 30 truck driver jobs there may be available 100 men whose outstanding civilian skill was truck driving. Since only 30 of these men can be assigned to truck driving duties, the simplest procedure would be to pick the 30 best drivers. The remaining 70, obviously, would be utilized in assignments other than truck driving. However, the solution is not always this easy. Suppose the demand for tank crewmen is more critical than the demand for truck drivers. To assign the 30 best civilian truck drivers to truck driving may result in sending poorer quality to tank crewman school—to train for the more critical specialty. Hence, it may be more desirable to

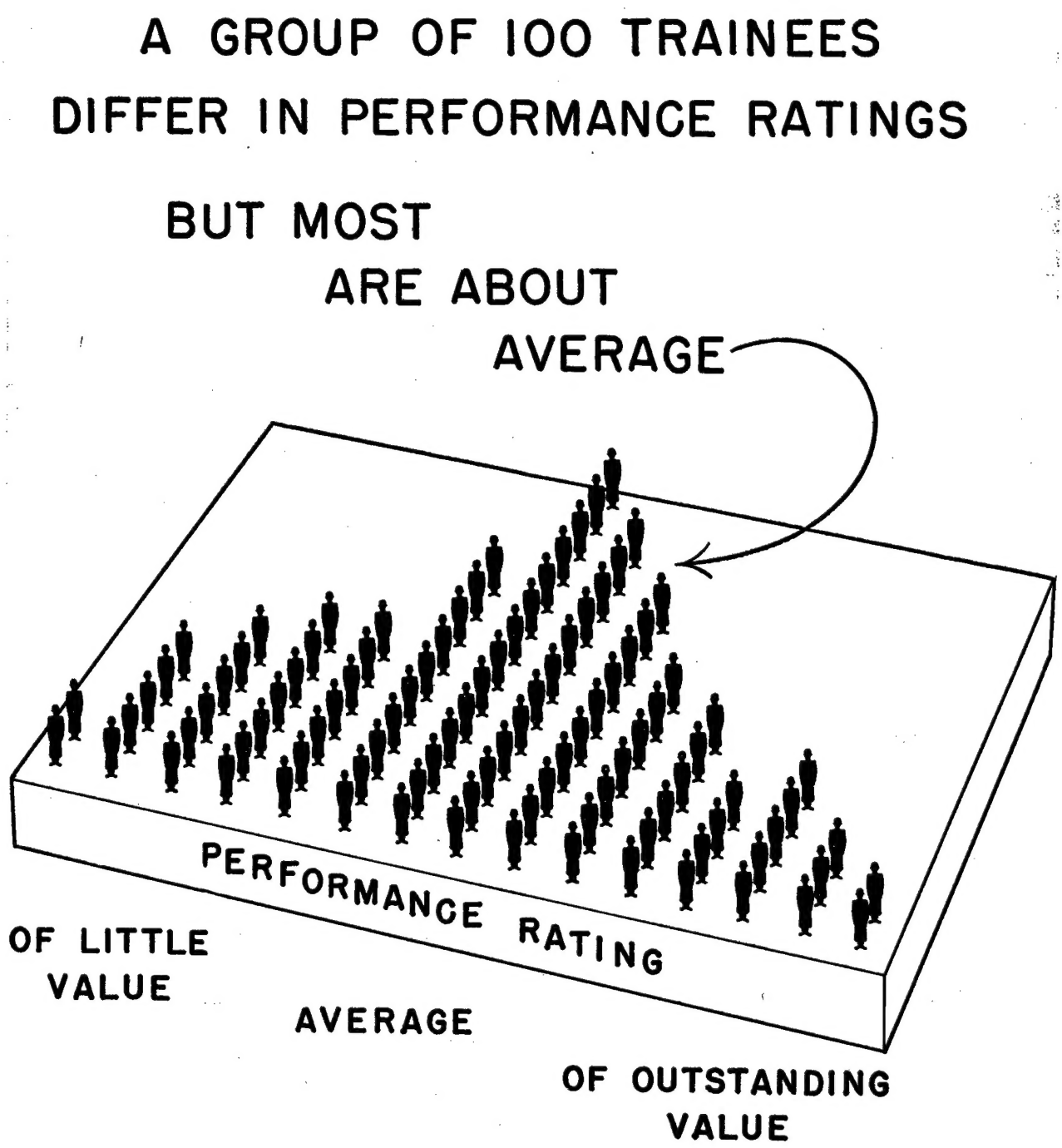


Figure 1. How men in a unit may differ in quality of performance.

assign the 30 best civilian drivers to tank crewman school.

b. Consider another example—a civilian lawyer may feel that by virtue of his training and experience he could serve best as a legal officer. This may be true as far as he is concerned. However, as far as the Army is concerned, there may be an oversupply of legal talent and an under-supply of artillery officers, a job which has no civilian counterpart. The problem then resolves itself into determining what the characteristics of good artillery officers are, and whether some of the lawyers possess those characteristics, so that they may be assigned to artillery training. A good lawyer may thus become a good artillery officer.

Section II. NEED FOR SCIENTIFIC PERSONNEL METHODS

6. Inadequacy of Personal Judgment

Some men are good judges of human nature. If there were enough good judges of human nature and if they had enough time to use their judgment critically, and if their judgments could be passed along to someone else without misunderstanding, there would be no need for other methods. Since none of these ideal conditions exist, it is necessary to rely upon the tools of personnel measurement, on aptitude tests, ratings, and the like.

a. Need for Objectivity. Some men may be able to "size up" a person and his qualifications with considerable accuracy. More often, however, such judgments are affected by prejudice. There may be insurmountable difficulties in passing along the judgment without misunderstanding. What one judge considers "good" performance may seem inadequate to another. Furthermore, personal judgments too frequently are not reasoned judgments. Too often they are no more than guesses—and uninformed guesses at that. There are some people who still believe that a man's ability and personality can be judged by his appearance. There is, however, abundant evidence that such judgments are generally too inaccurate for any important decisions to be based on them.

b. Need for More Efficient Methods. Personal judgment may serve as a personnel tool in Army units where a commander knows his subordinates, where he can decide personnel actions on an individual basis and observe how men work out in their assignments. However, this kind of classification is not common in the Army. Assignments to job or training more often are determined in classification or training centers where the information available about the man being assigned is limited to what has been obtained in the course of his initial classification and in subsequent actions such as performance ratings. It is essential, then, that this information provide the essential clues to a man's usefulness to the Army. It is not enough, however, to provide all the information essential to proper classification of a recruit. Such information must be in usable form. A classification officer, for example, who had at his disposal all the details of a soldier's personal and work history might be able to select an occupation which would be best for the man and, at the same time, work to the best interests of the Army. But it would be a difficult task and would mean consideration of a great number of nonessential facts. It is obvious that such a detailed study of every man and woman coming into the Army is neither feasible nor desirable. It would take too much time; there would be no objective guide to the soundness of the decision. Nor is it necessary; there are other methods which are as effective and even more effective—the methods of personnel measurement. The effectiveness of selection, classification, and promotion would be seriously reduced if personnel measurement tools were not available.

7. The Problem of Measurement

a. The problem of measurement in personnel psychology is fundamentally the same as in other scientific fields. It is the problem of defining the units in an unmistakable and useful way. What does the number 25 mean unless there is an added element such as "inches," "pounds," or "degrees"? Similarly, in personnel measurement, a number has no useful meaning unless the unit of measure is attached. Thus, a man may be assigned the number 25 because he gave correct answers to 25 test questions out of 26; he may also be assigned the number 25 because he gave correct answers to 25 test questions out of 100. Obviously, the number 25 does not mean the same thing in both instances; or stated another way, the unit needed is not just the number of problems correctly answered—the unit must also describe the number of problems he was required to answer.

b. Even this addition is not enough to give useful meaning to the number. How many problems should he be required to answer? Suppose two tests of 100 problems each are given to a group of men. Suppose, further, that on one of the tests perfect scores are made by a number of the men, but that on the other test, the best score ever made was, say, 65 right. Again, obviously, a score of 25 on the two tests does not mean the same thing, even though both tests had the same number of problems or items. It

does not mean the same thing because it does not represent the fact that one test was relatively easy (that is, perfect scores were made by a number of the men), whereas the other test was relatively difficult (that is, no one ever approached a perfect score).

c. Without attempting to be exhaustive at this point, only one further consideration will be raised. Even after all these definitions are attached to the number, there is still lacking one essential element of meaning. Is the number a "good" score or a "poor" score? Is the man who gets that numerical score a "good" man or a "poor" man? What, if anything, does the number reveal about the man's performance on other tests, the adequacy of his duty performance, his capacity for benefiting by further training, his promotability?

d. To sum it up, the problem of measurement in personnel psychology is not usefully solved until the numerical values used make clear how a man stands in comparison with other men and how well performance on the test predicts other performance. That is to say, standardization and validation are necessary to give useful meaning to the numerical values.

Section III. PERSONNEL MEASUREMENT IN PRACTICE

8. General

The primary purpose of all Army personnel evaluation instruments is to aid in measuring actual or potential job proficiency. Job proficiency does not mean just the performance of specific operations or movements. In some jobs, proficiency consists, to a large extent, of the ability to work with others as a member of a crew or team. On other jobs, leadership ability may be a prime requisite. Most combat jobs demand "courage under fire," and certain other jobs require a large amount of ability to understand people.

9. Progress in Personnel Measurement

a. Some human characteristics are easier to measure scientifically than others. For example, it is relatively easy to develop measures of ability to do the work at an Army technical school, but it is much more difficult to develop measures of "courage under fire." It is difficult to get a measure of "courage" not so much because courage cannot be measured, but because it usually cannot be studied scientifically where it counts most, in combat, for example. But even here a start is being made.

b. The greatest progress has been made in measuring physical traits, characteristics such as dexterity and intelligence, and skills and knowledges of many kinds. Considerable progress has been made in measuring personality characteristics such as those involved in leadership, although much work remains to be done to obtain more exact and dependable methods of measurement. An important source of difficulty is that such characteristics are very complex and may depend on specific situations. That is, a person who is a good leader in one kind of situation may be a poor leader in another. It is apparently not useful to obtain a measure of "leadership"—it appears necessary to add "leadership for what, in which circumstances, and of which kinds of men."

10. Calculated Risks in Personnel Measurement

No personnel measuring instrument is perfect; that is why continuing research is needed. A major objective of personnel research in the Army is to strive continuously to improve the test, the rating scales, the personality inventories and the other instruments used so that the amount of "error" in the measurements will be a minimum. What is of equal importance is knowing the limits of accuracy with which an instrument may be safely used, and separating what it is feasible to measure from what it is not feasible to measure. This knowledge is important not only for the proper interpretation of particular scores but also for establishing policies governing personnel action. Should additional personnel information be sought? What follow-up is needed? Who should make the final decisions? To answer such questions it is necessary to know the limits of accuracy of the measuring devices. Thus, "calculated risks" are involved in personnel actions as well

as in tactical operations. Research in personnel measurement provides a basis for calculating the risks in personnel actions.

11. Personnel Research as a Practical Approach

In order to furnish a basis for calculating risks in personnel actions involving personnel tests and other instruments, it is not enough to develop a theory or to speculate as to an instrument's value. It is necessary whenever possible to "test the test." How consistent is the instrument every time it is applied—that is, what is its reliability? And even if it is highly consistent, what is it really measuring? Not what someone thinks it measures, but what it really measures—that is, what is its validity for the use intended? The answers to these vital questions require a practical approach—field testing whenever possible. It is not an easy approach, for field testing requires a yardstick or criterion with which the results of administering the instruments can be compared, and such criteria are not easily arrived at. Problems concerning the use of yardsticks or criteria for determining the effectiveness of personnel measuring instruments are so important that a whole chapter of this manual is devoted to them (ch. 3).

Section IV. THE SCOPE OF PERSONNEL RESEARCH

12. Classification

a. Personnel research in the Army is directed primarily at improving methods for selecting and evaluating personnel as individuals. The psychological and physical characteristics of a man must be measured before the Army knows whether he is soldier material. It is necessary to measure these characteristics at various stages of a soldier's career to determine what assignments he can be expected to perform satisfactorily and what special training it is profitable to give him. It is necessary to measure the outcomes of the training given him. Throughout his military career his abilities must be reviewed and evaluated. As he acquires new military skills through training and experience or as the changing needs of the Army demand, his assignment is subject to revision.

b. Personnel research is also directed at improving methods for selecting and evaluating personnel as teams. It is necessary to fit together men with different characteristics to permit effective accomplishment of the mission of the team. The effectiveness of the team must be evaluated to determine what missions it can be expected to accomplish, what special training it is ready for, how much it has profited by the training given it, and how it has performed during tactical operations.

c. It is the business of personnel research to develop tools for selecting and evaluating personnel, as individuals and as teams, so that assignments and changes in assignment may have a sound basis.

13. Personnel Research and its Relation to Other Areas of Research and Policy in Human Resources

a. Personnel research in the Army is concerned primarily with the scientific development of devices to select and evaluate personnel. As a research activity, personnel research may have important relation to other areas of research in human resources, such as training methods, psychophysiology, personnel management, and statistical methods. As a tool to improve the Army's utilization of manpower, personnel research may contribute to the making of personnel policy and is often the key to its execution.

b. Some examples may be given of the relations between personnel research and other problems involving military personnel. Suppose that a service school wishes to improve its tests and its grading system. Study may indicate that the tests and the grading system are adequate. Question may then arise as to the adequacy of the methods of instruction. Accordingly, research may be undertaken to compare several methods of instruction, and personnel measurement may be used to evaluate the effectiveness of the various methods. Or a question may be asked: What is the relationship between qualification in arms and over-all value as an infantryman? In searching for the an-

swer to this question, another question may arise—Are the current methods of weapons training in need of improvement? This may lead to still another question—Are the weapons designed to fit the men's abilities and thus permit the effective use of the weapons?

c. Suppose that research is under way to develop better tests of vision as part of an effort to determine how visual performance affects over-all performance in various jobs. Before such tests can be developed, it may be necessary to study the difference between, say, night vision and day vision. Or it may be desired to develop tests which will detect likelihood of accidents among motor pool drivers. However, before the tests are developed, research may be needed to clarify what is meant by "accidents" and to improve methods of recording accidents. Or, for a final example, when the minimum qualifying score on the test for induction into the Army was to be set, a basis for the decision was provided by personnel research findings. The score as set reflected the best balance that could be struck between the number of men of marginal usefulness available to the Army and the number of such men who could be absorbed into the Army's job structure.

Section V. SUMMARY

14. The Purpose of Personnel Research

a. Army personnel research is concerned with discovering techniques that will facilitate the effective utilization of its manpower. This requires—

(1) Analysis of the psychological requirements for each job.

(2) Analysis of men in terms of—

(*a*) Individual differences in abilities.

(*b*) Their civilian specialties in relation to the manpower needs of the Army.

b. The size and scope of the Army requires the application of scientific personnel methods so as to increase efficiency in its personnel actions. Objective indicators of performance should replace personal judgment where possible, and effort should be directed at improving and systematizing personal judgment where needed.

c. The Army emphasizes a practical approach in its personnel research. Its purpose in using personnel evaluation instruments is to aid in measuring actual or potential job proficiency. Achievement of this purpose requires knowledge of the limits of accuracy of the instruments and is conditioned by the following facts:

(1) Some human characteristics are harder to measure than others.

(2) Validating measuring instruments requires considerable time and effort.

d. Personnel research is one area of human resources research. The products of personnel research are used by the Army to aid in the classification and utilization of personnel, as individuals and as teams. These products may also be used in other areas of human resources research, such as the evaluation of various training methods. Personnel research may be used in the solution of other problems of personnel management and in the establishment and execution of personnel policy.

CHAPTER 2

HOW THE ARMY DEVELOPS ITS PERSONNEL MEASURING INSTRUMENTS

Section I. MAJOR CONSIDERATIONS

15. General

a. Most Army personnel measuring instruments are designed to be used for a particular personnel program. To apply an instrument to a personnel program other than the one it was designed for is risky. The nature of the population taking the test may make considerable difference in the usefulness of the test. For example, an instrument developed for inductees is not necessarily applicable to trainees or combat veterans even though the psychological characteristics to be measured appear to be the same. A test to measure how much men in basic training have learned about the functioning of a particular weapon might well be too easy for combat veterans. If it is used with combat veterans, it will not be possible to determine which of them know most about the weapon, since they will answer most of the questions correctly. Another test with more difficult questions is necessary.

b. The principles and practices followed in the development of personnel measuring instruments are described in this chapter. For simplicity's sake, the discussion is oriented toward the development of tests. It will be understood that the discussion applies, in general, to other types of personnel instruments as well.

16. Types of Instruments

a. The Personnel Measuring Instruments Used by the Army Are of Several Different Forms. Distinctions are often made between tests on the one hand, and rating scales, questionnaires, and interviews on the other. The term "test" is usually used for an instrument that requires answers which are either right or wrong. By way of contrast, the other types of instruments do not have right and wrong answers but rather yield indication of the degree to which a characteristic is possessed. In either case, it is important to determine the significance of the scores, whether they be based on the number of right and wrong answers or by the possession or absence of particular characteristics. As stated in chapter 1, it is necessary that instruments be valid if they are to be used effectively.

b. Personnel Measuring Instruments May Be Classified According to What They Are Intended to Measure. Some are intended to measure knowledge of a job or subject matter; others are intended to measure ability to acquire such knowledge. Both types of instruments are commonly called tests.

(1) Some instruments (ratings) are intended to evaluate a man's performance, usually in terms of his value to his organization; others (self-description forms) provide a means for the man to describe his past history, his likes and dislikes.

(2) A standard interview is an instrument used to evaluate how a man acts under prescribed circumstances. Sometimes an interview is used to find out what a man knows, although usually this can be done better with a suitable test or records (ch. 8). For special purposes, an interview, usually called a "stress" interview, may be used to evaluate the actions of a man under severe pressure. This type of interview is not widely used in the Army.

(3) Some measuring instruments are directed at obtaining knowledge of a

man's attitudes and beliefs. Such instruments are usually not involved in personnel actions and hence are not discussed in detail in this manual.

c. Personnel Measuring Instruments May Be Classified in Other Ways. Some are administered to one man at a time; others—and in the Army, most—are administered to large groups. Some (most) require the use of language, usually English; others require only a minimum of language. Some (most) are of the familiar "paper-and-pencil" type; others require performance of a work sample. Some instruments are called aptitude tests; others, achievement tests.

17. Importance of Understanding how Personnel Measuring Instruments are Constructed

All classification personnel in the classification system and all officers exercising a command function in regard to classification or assignment need a thorough understanding of the nature of Army personnel instruments and the proper interpretation and use of scores. Such understanding can be aided by a brief description of the principles and practices of constructing instruments.

18. Tests are Designed to meet Specific Army Needs

a. Defining the Army Need. The first step in test making is to study the classification problem to be solved. The particular need of the Army must be clearly defined both in terms of the psychological requirements of the job or course of training and in terms of the number of men needed and the apparent supply. As a rule, the difficult and time-consuming process of instrument construction is warranted only when the problem is urgent and important. If a simple interview, a survey of past experience, or an examination of personnel records will do just as well, or if the number of men and jobs involved is small, the construction of a special instrument is not warranted.

b. Finding the Characteristics to Measure. Study of the problem also brings to light the characteristics which are related to successful performance. It is necessary to find characteristics which are possessed in high degree by most (if not all) of the men who have demonstrated their ability in a particular course or assignment and are not possessed by those who have failed. By measuring the amount of these traits possessed by untried men, it is possible to predict their performance on the job. In some cases, it is a simple matter to find characteristics highly correlated with success—successful carpenters know such things as the proper use of various tools and the various types of joints. Obviously, these are the skills and knowledges that should be measured in order to select the Army's carpenters. In other cases, the particular traits must be discovered through systematic trial made after the general purpose of the test is decided upon. For example, it was necessary to measure "cryptography aptitude" in order to predict which men would be most likely to pass a course in cryptography. It took considerable research to find which particular traits made up cryptography aptitude and to choose those most highly correlated with successful performance in the course. It is essential to select for measurement those traits which are actually to be found in the men themselves and not in the circumstances of a training course or assignment. If, for example, men fail a course because it is heavily loaded with irrelevant material, there is no use in devising a test to discover the most likely candidates for the course.

19. Suiting the Instrument to Its Purpose

The use to which an instrument will be put is one of the principal factors which determines the kind of test to be developed and the form it will be given.

a. Achievement Tests—Instruments Intended To Find the Men Who Possess the Skill and Knowledge Necessary for Successful Performance in a Particular Assignment. Thus for example, the Basic Military Subjects Test is designed to measure the amount of job knowledge that a soldier has acquired as a result of basic training. If it is known how much of this knowledge is necessary for satisfactory performance in later duty assignments, then the scores on this test may be used as aids in estimating future level of performance. Achievement tests most often are used in selecting men for direct

assignment to some tactical or service organization. They contribute to the prediction of job proficiency by measuring how much knowledge or skill the man has already acquired.

b. Aptitude Tests—Instruments Designed to Estimate in Advance Probable Performance. They may predict which men are most likely to perform successfully in a training course or duty assignment. They predict success by measuring the degree to which examinees possess certain critical traits found in men who have been successful in the particular training program or duty assignment for which selection is being made. They are useful in selecting from among a mixed lot of available soldiers the most promising trainees for a particular course. They are particularly useful with the recruit whose job experience has been limited and whose potentialities for various kinds of available training are unknown. The fact that certain men get high scores on such tests would have little significance were it not for the fact that research studies have shown that such men are likely to benefit more by subsequent training for certain jobs than are men who get low scores.

c. Whether a Personnel Instrument Is Considered as Measuring Aptitude or Achievement Depends Mainly Upon the Purpose of the Instrument. Hence it is often difficult to differentiate between measures of aptitude and measures of achievement on the basis of test content alone. Aptitude may be measured by an achievement test, provided that the possession of certain knowledge and skill indicates aptitude for acquiring other related knowledge and skill. For example, a radio achievement test can be used in the selection of men to be trained as radar technicians, since research studies indicate that men who have acquired a knowledge of radio have aptitude for the more highly specialized radar.

20. Form of Test

Two principal factors determine the form in which a test is cast—the purpose to be served and practicability. If a test is to be used on illiterates, obviously it must be in a form which minimizes the amount of language required. If, for example, a test is intended to show how a man does a job requiring physical handling of equipment and materials, a paper-and-pencil test of his knowledge of the job is not appropriate. Furthermore, a test is a field instrument and must be designed to serve efficiently under conditions likely to be found in the field. It must

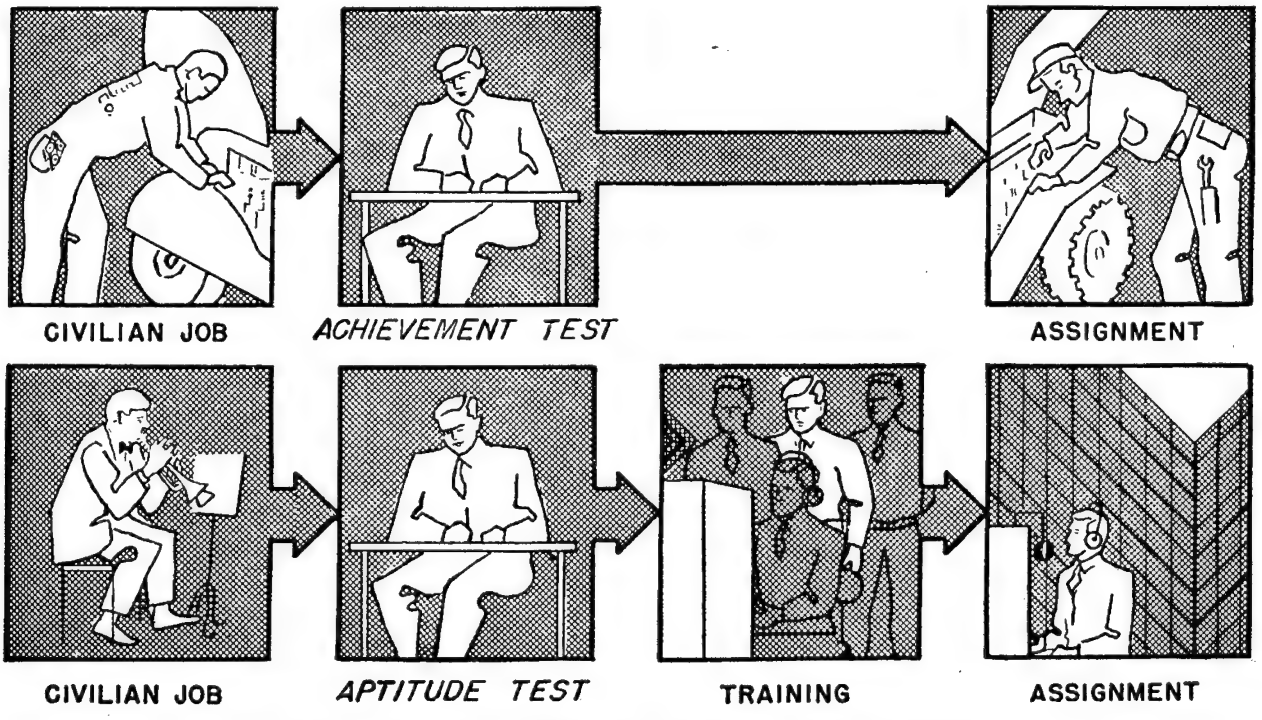


Figure 2. Aptitude and achievement tests serve different purposes.

be adaptable to Army necessities which limit the facilities available and the time that may be allotted to testing.

a. Verbal Test. A verbal test is one in which the examinee is required to talk, write, or mark correct responses stated in language. It is not limited to tests of vocabulary and reading. Most Army tests are verbal tests. The type commonly used is a paper-and-pencil test administered and scored in a short time and with great efficiency by men who need not be highly trained specialists. Conditions of administration can be made uniform, and highly objective scoring methods may be employed. A wide range of ability and knowledge may be sampled in a relatively short time. When a large number of men are to be tested, or highly trained personnel are not available to administer tests, a verbal test is likely to be the most practical, unless it is clearly unsuited to the purpose at hand.

b. Non-Language Test. Where the examinees are illiterate or do not understand English, it is necessary to use pictures and diagrams instead of verbally-stated problems, and to substitute demonstration for verbal instruction in administering the test. The Non-Language Test 2abc, administered to non-English speaking aliens enlisted by the United States Army in foreign countries, is an example of this type of test.

c. Performance Test. A performance test is one in which the examinee is required to manipulate objects, or make practical application of knowledge. It is the most efficient form to use when it is necessary to observe how the examinee does the job as well as his ability to do it, or when it is essential to measure ability to do a job rather than knowledge about the job, although the two are usually correlated with each other. The practicality of performance tests is limited in that they are time-consuming, expensive, usually require highly trained personnel for their administration, and are frequently difficult to score. The Machinist Performance Test, an example of this type of test, is discussed in chapter 7 as a work-sample achievement test.

d. Group Tests. Verbal (paper-and-pencil) tests are widely employed by the Army because they are best adapted to group testing. The large number of men to be classified and the pressure of time make it necessary to test in groups whenever possible. Performance tests, by their very nature, are better adapted to the testing of men one at a time. It is sometimes possible, however, to devise a performance test to be given to groups.

e. Individual Tests. Individual tests are administered to one man at a time. They may be of either the verbal or performance form, depending upon the particular purpose to be achieved, or a test may involve both performance and verbal responses. Individual tests are constructed to accomplish the following purposes for which group tests are not suited:

(1) Screening and classifying men whose language deficiencies or other personal characteristics do not enable them to demonstrate their abilities adequately on a group test.

(2) Testing aptitudes or proficiencies which can best be revealed through actual work samples or manipulation of materials.

Sometimes, an individual test is administered to a person to help determine from his behavior while taking the test whether his test score is consistent with other information available about him.

21. Multiple-Choice Items

Two steps in constructing the test thus far have been considered. The traits and characteristics to be measured have been decided upon, and the form which the test is to take has been determined. The next step is to determine in which form the questions or problems (that is, the test items) should be stated to reveal the desired information. They may be brief statements requiring essay answers. Or they may be incomplete sentences or paragraphs requiring insertion of the correct terms. Other forms are possible. For a variety of reasons, the multiple-choice item has been adopted by the Army for large-scale classification testing.

a. Description of Multiple-Choice Items. In performance tests, the items usually are problems involving cards, blocks, tools, equipment, etc. In paper-and-pencil tests, the items are generally multiple-choice questions or statements. Multiple-choice items present the examinee with several (usually four or five) an-

swers to a question. The problem then is to indicate the correct answer. Examples—

(1) Boston is the capital of—
  (*a*) Maine.
  (*b*) Montana.
  (*c*) Massachusetts.
  (*d*) Minnesota.

(2) In the Diesel engine, the gas mixture in the cylinder is ignited by the—
  (*a*) Spark.
  (*b*) Heat generated by compression.
  (*c*) Ignition system.
  (*d*) Firing order of the cylinders.

b. Advantages of Multiple - Choice Items. Multiple-choice items are preferred for most classification testing for several reasons. Scoring is more objective because the right answer is already set down and is not subject to the varying judgments of testing personnel. The examinee has only to recognize the answer, and is not burdened by having to search for it in his mind and then phrase it in his own way. Because the examinee does not have to write but is merely required to check the correct answer, he can cover many multiple-choice items in a short time. Multiple-choice items can be scored by machine, (as described in ch. 12) which *increases accuracy* and *saves time.*

22. Length of Test

a. General. To include all the items pertinent to a given trait would make the test absurdly and inefficiently long. To include too few items pertinent to a given trait would result in a test which would not measure the trait reliably. The principle which governs the number of items used in the test, and therefore the length of the test, is that there must be enough to show the degree to which each examinee possesses the trait, and to show this in a measurable fashion so that men can be compared with one another in terms of the trait.

b. Classification Testing Employs the Same Sampling Principles Followed in Other Fields of Measurement. In grading a carload of wheat, for example, it is not practicable to examine the whole lot in order to compute the percentage of high-quality grain, of chaff, and of foreign materials. The examiner instead gathers samples, analyzes them, and assumes that the characteristics of the whole carload are the same as those of the sample which he tested. But he would be extremely naive if he took all his samples from the top of the car. An unscrupulous vendor could easily have filled the car with an inferior grade of wheat and placed a thin layer of first class stock on top. The sample taken from this top layer would not be representative of the whole, and the measurement based on this sample would be an exceedingly inaccurate indication of the quality of the entire lot. Aware of all the pitfalls of careless sampling, and wishing his sample to be representative of the whole carload, the examiner collects a number of smaller samples—from the top and bottom of the car, from different depths, from each end, and from the middle. The more samples he collects, the more accurate his grading. The car may contain a concentration of inferior wheat, constituting a very small fraction of the total amount. If the examiner takes only five samples, and happens to take one of them from this small concentration, the inferior grade will comprise one-fifth of his total sample, and conclusions based on it will not apply to the whole carload. So it is with testing. Enough items should be used to sample adequately all important parts of the area which the test is to measure. Further discussion of sampling is presented in chapter 5.

c. Practical Considerations Limit the Length of the Test. In practice, the test-maker includes as many items as are necessary to make the test sufficiently accurate for sound classification, but keeps it short enough to be practical under field conditions and within the limits of fatigue and boredom of the average man.

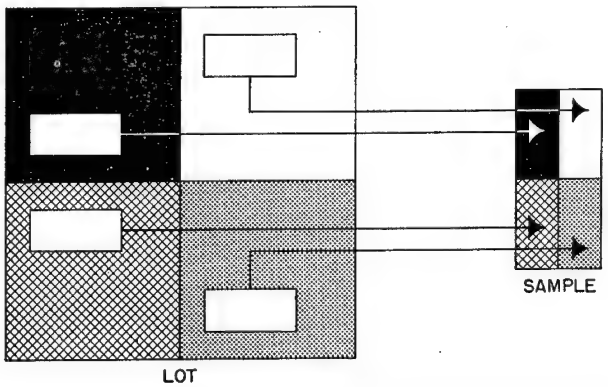


Figure 3. A good sample comes from all parts of the lot.

23. Time Limits

Time limits are extremely important because they help to determine the qualities measured by a set of items. A test made up of items which are equal in difficulty measures speed if the time limit is so short that no one can finish all the items. A test in which the items get successively harder and harder measures power if examinees are given all the time they need to complete as many items as they possibly can. Most Army tests measure both power and speed. The Army must consider both how well and how fast a man can be expected to perform in a given assignment. The Officer Candidate Test is a good example. It is made up of questions and problems which become harder as the examinee forges through them. Only the man who is both able and quick can complete most of the 70 items in the 45 minutes allowed for the test.

24. Other Considerations in Planning Tests

The nature of the particular instrument may introduce special problems. The decision as to a speed or power test may be important for achievement testing; it is not so important for such instruments as interviews and self-description forms (chs. 8 and 9).

Section II. PREPARING THE TEST

25. General

After basic decisions have been made concerning the traits to be measured, the general form of the test and the type of test items, the next task is to prepare an experimental form of the test. Preparation of the experimental form includes the actual writing of the items and preparation of a scoring key and manual. This step is only the beginning. Only after a test has been subjected to field trial and a useful scale of measurement established can the test be considered as properly developed.

26. Construction of Test Items

Prior to the construction of items, it is necessary to study the performance of men in the training courses or duty assignments for which the test is to be used. This study may take the form of job analysis involving direct observation of the men. It may take the form of examination of training curricula, consultation with experienced men, or examination of official job descriptions. Such information is valuable in furnishing detailed ideas for the content of items. Then a large number of questions or problems, closely related to the trait to be measured, are assembled. When appropriate, the writer of the items consults a subject-matter specialist to aid in keeping the subject matter of the items on a sound basis. Each item is carefully checked to make sure that it is logically and clearly stated and that it seems likely to produce an answer that will indicate something about the trait which is being measured. Finally, all the items are reviewed by a group of experts who may recommend final changes in form, phrasing, or content. The items which remain are then collected into an experimental booklet. Directions for administration and scoring are prepared for use during an experimental try-out.

27. Preparing Directions

a. The Directions Which Accompany Each Set of Test Items Constitute a Statement of the Conditions Under Which the Test Was "Calibrated" or Standardized. Great pains are taken to make these directions complete and clear. Only by following them can the standard conditions—the conditions under which the test was standardized—be repeated (ch. 11). Unless these same conditions prevail, a test gives results as unreliable and misleading as a thermometer reading taken when the patient has a mouth full of ice.

b. Two Sets of Directions Are Prepared for Each Test: One for the Examiner and One for the Examinee.

(1) Instructions and suggestions to the examiner are included in the manual that accompanies the test whenever it is administered. They indicate the general conditions under which the test should be given, list the materials required to give it, and state time limits for the parts and for the whole test.

They also suggest introductory remarks that should precede and set the stage for the administration of the test, as well as recommended answers to questions that commonly arise during the testing session.

(2) The second set of directions provides the specific instructions to the examinees. These are printed as part of the test booklet itself to insure that instructions will be the same for every administration of the test. Their purpose is to make certain that each examinee understands just what he is expected to do. They touch upon such details as the advisability of guessing when not absolutely sure of the answer, the amount of time that will be allowed for the test, the relative importance of working for speed as against working for accuracy. And they give precise and detailed explanations, along with demonstration and practice items, of the correct manner of indicating answers. Since all of these directions are such a vital part of the test, they are prepared by experienced test-makers and subjected to independent check for completeness and clarity.

28. Preparation of Scoring Directions

a. Placing the Items and Answers in Proper Order. The final step in making an experimental model of a new test is to work out the proper technique for scoring. The items are first put in the order which it is believed will yield the best results. The position of the right answers is adjusted so that they fall in random order. That is, the right answers are so located from one item to the next that the examinee will not be able to "outguess" the test by discovering a particular pattern of right answers. A scoring key indicates the position of the right answer on the answer sheet. The correct answer is placed in the corresponding position in the test booklet. The incorrect choices or alternatives are so arranged as to eliminate any clues that the examinee might get from the order of the alternatives. A final check is made to be sure that the right answers shown on the key and the correct alternatives in the test questions correspond.

b. The Scoring Formula.

(1) *Correction for guessing.* In some types of tests guessing may give a man a higher rating than his abilities warrant. The following example shows why the scoring formulas described in detail in chapter 12 are sometimes applied. If examinees select one of four alternative answers to a multiple-choice item by pure guess, their chances of selecting the correct alternative are one in four. In a large number of such guesses, they are likely to be wrong three times for every time they guess right. On a 100-item test, for example, they will usually obtain about 25 right choices by answering in this fashion. Since they answered one item right for every three they answered wrong, an estimate can be obtained of what their scores would be if they had not guessed by subtracting from the number right one third of the number wrong.

(2) *Is correction for guessing useful?* The use of the correction formula is based on the logic of chance. In practice, however, guesses are seldom completely blind. An examinee may get an item right by knowing which alternative is correct or by knowing that the other three are wrong. Likewise, if he knows that two are wrong, he will have to guess only between the remaining two and will, therefore, stand a better chance of picking the correct one. Any error that results from the application of the correction formula will always be in favor of the examinee who utilizes such judicious "guessing". However, there will be other, more cautious examinees who may know just as much but who will never put down an uncertain choice if they are to be "penalized for wrong answers." It has been found that the correction formula is not sufficiently helpful to apply in every situation. It is therefore used only where appropriate.

Section III. TRYING OUT THE TEST

29. General

After planning and constructing the experimental model of the test, the next step is its trial run. The aims of the trial and analysis of the new test are, in general, the same as those for any other trial run. The pilot model of a new gun is tested to make sure that it will shoot accurately and consistently and according to its specifications. Similarly, the pilot model of a new test is given a trial to make certain that it will measure specified characteristics with sufficient accuracy to permit construction of an operational test that will be effective in classification of soldiers.

30. Field Trials and Analysis

In the experimental form of a test, there are many more items than will be used in the finished product. This is necessary because, on the basis of the findings obtained in field studies, certain of these items will be rejected. In addition to making certain that enough items are available to allow for dropping unsatisfactory ones, certain other considerations are important.

a. The Population Used in the Field Trial.

(1) Ideally, the men to whom the experimental test is administered should be representative of the men who will take the finally developed test. A test to be used with inductees should be tried out on a group of men representative of inductees, not on a group of men who have completed advanced schooling. A test to be used for selecting enlisted men for officer candidate school should be tried out on a group of enlisted men otherwise eligible, not on men already enrolled in officer candidate school. In brief, the experimental sample should truly represent the operational population from which selection is to be made (ch. 5).

(2) Unfortunately, this ideal can only be approximated and too often the extent to which the experimental sample deviates from the operational population is not known. For one thing, the operational population may vary markedly from time to time so that its characteristics are not stable. For another, the essential *criterion* data (ch. 3) needed to determine the effectiveness of the test may not be available for all the men who will take the operational test. One example will illustrate this point. A test which is intended for use in selecting enlisted men for a school will obviously lack the school data for the men who are not accepted. Thus, in estimating the effectiveness of the test in predicting how well men will perform at the school, the estimate will contain a certain amount of error arising from the fact that the men in the school are only the better men—that is, they represent a range of ability restricted to the higher levels.

(3) Various practical attempts are made to allow for the possible lack of representativeness of the experimental samples. Statistical methods are available for correcting for restriction in range when such correction is justifiable. These methods are beyond the scope of this manual and will not be described here. Another procedure is that of conducting follow-up studies of the effectiveness of the test on succeeding groups of men. This method has the advantage of taking into account possible variations in the range of abilities represented by the men from time to time. Stated another way, the effectiveness of a test is not determined once and for all; the effectiveness is examined over a period of time.

b. Difficulty Index of Items. Another consideration in the field trial of an experimental test is the determination of the difficulty of each of the items. Difficulty is not determined by the subjective estimate of the test-maker, nor by the subject matter expert, nor by any opinion that the item "should be easy for anyone claiming to be familiar with the content of the test." The difficulty of a test item is determined from

actual try-out: it is defined as the proportion of men in the group who actually answer the item correctly.

(1) *How item difficulty is expressed.* The estimate of item difficulty from the trial run, therefore, involves the simple but tedious task of counting, for each item, the number of examinees who answered it correctly and computing percentage of the total number of cases. Difficulty is thus expressed in percentage form; a difficulty of 70, for instance, means that 70 percent of the group answered the question correctly. This would be a relatively easy item. An item having a difficulty of 28, that is, an item answered correctly by only 28 percent of the group, is relatively hard. (It will be noticed that a low percentage means difficult and a high percentage means easy.)

(2) *How item difficulty is used in item selection.* In tests involved in personnel actions, there is little point in including a large number of very difficult or very easy items. If some of the items are so difficult that no one answers them correctly, then all that is accomplished is subtracting a constant from all the scores without affecting the ranking of the men on the test. Similarly, including easy items that every one answers correctly means merely that a constant is added to every man's score. This point is elaborated in paragraph 31.

c. Internal Consistency Index of the Items. The internal consistency index of an item refers to the relation between it and the other items in the test. If, for example, the men who answer a particular item correctly tend to answer most of the other items correctly, while those who answer it incorrectly also answer most of the other items incorrectly, then that item may be considered to be consistent with the other items. It will have a high internal consistency index. This index must be used with caution. If items with high internal consistency indexes are added to a test, the test as a whole will measure more consistently whatever the original items measured. However, the validity of the test, or the degree to which it agrees with some particular criterion, is not thereby improved unless it is known that the original items were valid and that they all measured the same thing (that is, that the test was "pure"). The internal consistency index of an item gives information about the external validity of the item only when the test as a whole is valid in predicting an external criterion.

d. Validity of Items. There is one other characteristic of an item that it is necessary to know. Is it answered correctly by good men (high on the trait being measured) and incorrectly by poor men (low on the trait measured)? If half of the men who answer an item correctly are known to be competent and the other half are known to be incompetent, that item is useless in distinguishing the good from the poor. In a well constructed achievement test that is intended to determine, for example, what men have learned in a particular course, the validity of an item rests upon whether the content of that item was adequately covered in the course. In such cases, the item is said to have "face validity" (ch. 5).

(1) However, when such a test is used to indicate how well the men will perform on the job after completion of the course, validity of items in the test must be established in terms of job success. Suppose that successful job performance requires not only knowledge of course content but ability to work with others, ability to adapt materials to the purpose at hand, ability to solve problems not covered in the course, ability to lead, and so on. In such cases, it is necessary to discover if men who are good on the job answer the item correctly and men who are poor on the job answer it incorrectly. If this is the case, the item is said to have "high validity." If both competent and incompetent men on the job answer an item similarly, the item has low validity, and there is no point in including it in the instrument.

(2) The determination of the validity of items in such instruments as self-description forms is especially impor-

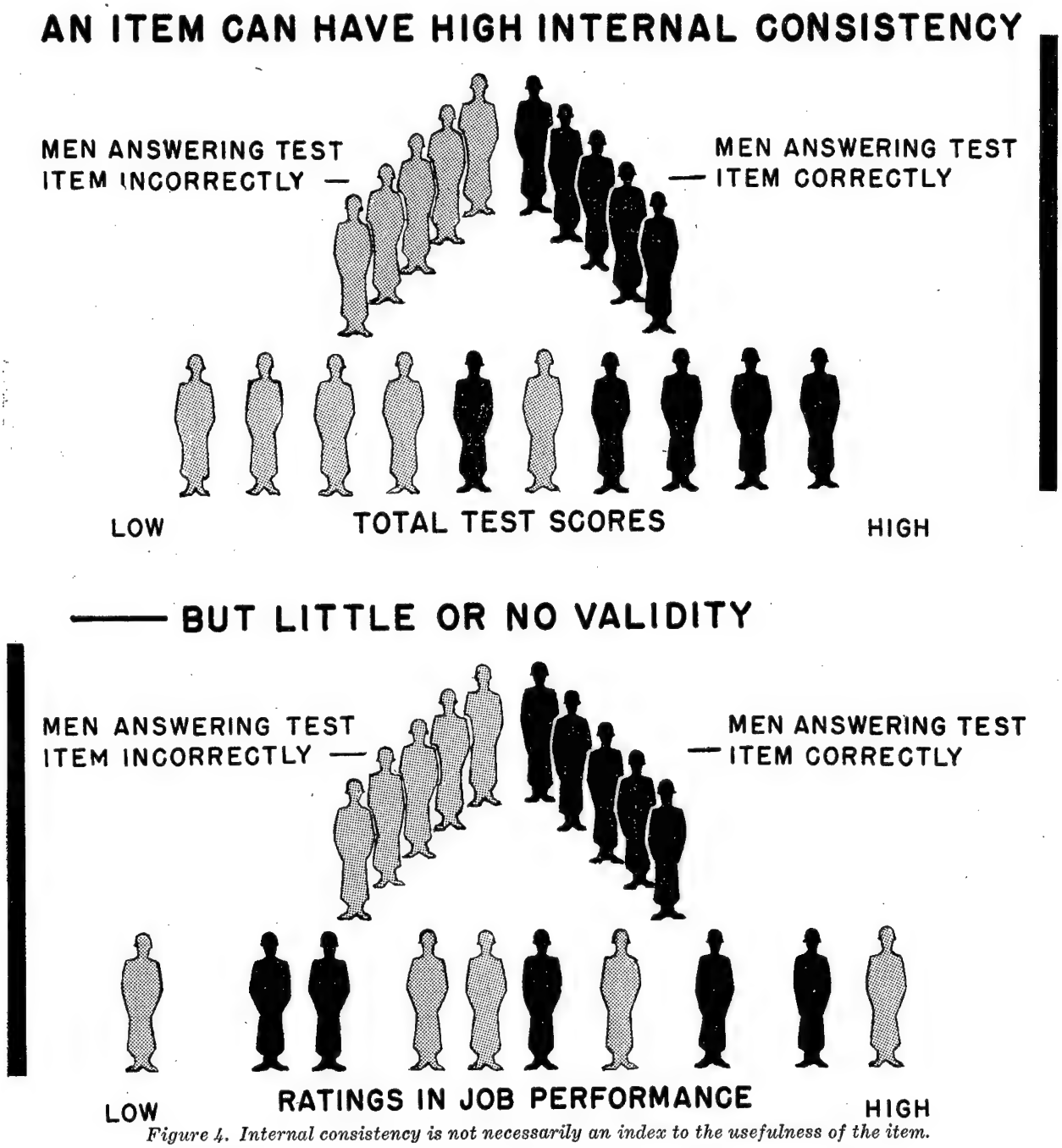


Figure 4. Internal consistency is not necessarily an index to the usefulness of the item.

tant since such items do not have correct or incorrect answers but rather indicate the degree to which a particular characteristic is present. It is still necessary to determine whether men who are good in their job performance tend to select one answer and those who are poor select another. Where this is not true, the item is useless for discriminating between good and poor men.

31. Selection of Test Items

The field studies of the experimental form will have furnished data on the three main characteristics of each item—its validity, its difficulty, and its internal consistency. The evidence is now ready for use in selecting the best of all the items tried out.

a. Validity. The first and most important consideration is the validity of the item. The items are examined and those with little or no

validity are eliminated, regardless of their other characteristics. There is little point to loading a test with items which have little to do with what the test is supposed to measure. They may be valid for other purposes, but if not valid for the purpose of this test, they add nothing to this test.

b. Difficulty. Once the items have been selected on the basis of their validity, they are examined to make certain that their range of difficulty is suited to the purpose of the test.

(1) If this purpose is to make the most efficient division of the population into a high and low group with reference to the trait being measured, the difficulties of the items selected should cluster around the division point. More specifically, if it is desired to qualify the top 30 percent of a population for specialist training or assignment, then the difficulties of the items selected should cluster around 30 percent (items answered correctly by 30 percent of the population).

(2) If—as is usually the case—it is desired to grade the whole population from highest to lowest with reference to a trait rather than merely to divide it into two groups, the difficulties of the selected items should be spread over most of this range. In most tests the item difficulties will be fairly evenly distributed over the range from 30 percent to 70 percent.

(3) In order to obtain the desired range of item difficulty, it may be necessary to accept some items with somewhat lower validity than other items have. However, this compromise is not extended to the point where items with no validity are included in the test to round out the range of difficulty.

c. Internal Consistency Index. As stated in paragraph 30, this characteristic must be used with caution. If the test is supposed to measure a specific ability, such as ability to manipulate numbers, then the items which are selected should have high internal consistency indexes. However, if the test is supposed to cover a broad area of abilities, such as understanding verbal statements or ability to persevere in a disagreeable task, then high internal consistency indexes may be of little value.

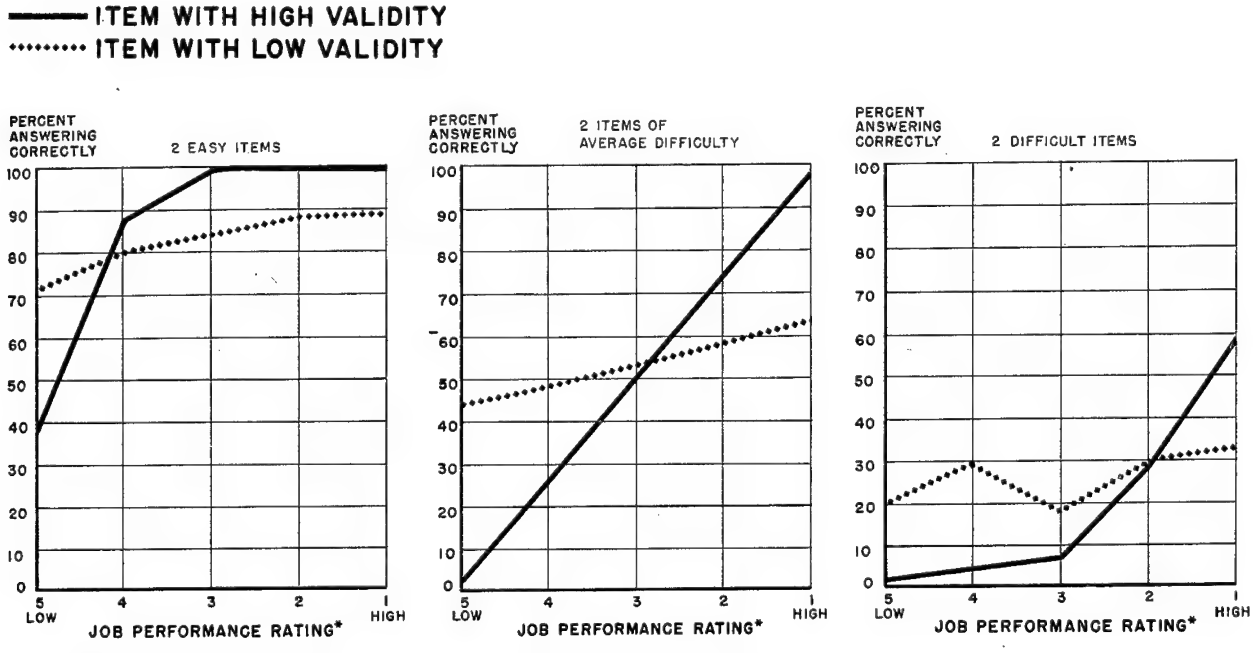


Figure 5. Illustrations of high validity and low validity for items at three levels of difficulty.

32. Final Form of Test

When the best items for the purpose of the test have been selected on the basis of data from the trial run, the test is in its final form. It is then important to establish the validity and reliability of the finished product (see ch. 5 on validity, ch. 4 on reliability). The test is administered to another group of men, selected so that they will be representative as far as possible of the population for which the test was designed (see ch. 5 on sampling). The final directions, time limits, and scoring procedures are used at this time. Scores of men tested are compared with measures of job or training success, and the validity of the final test is estimated. Reliability is also computed from these data. This whole process of checking the usefulness of the final form of a test on another and similar group of men is called *cross-validation.* It is an important step in test construction, a step which aids in estimating the value of the test under operating conditions. Cross-validation of a test may be accomplished more than once, as the Army population for which the test was designed changes.

Section IV. ESTABLISHING THE SCALE OF MEASUREMENT

33. Purpose of Standardization

a. Upon completion of the item selection, the test is a finished product in the sense that it has been made as accurate and dependable as possible. But measurements made with it will still be in terms of "raw" scores, that is, the number of questions answered correctly, or the number right minus a fraction of the number wrong, or, in the case of instruments which indicate the degree to which certain characteristics are possessed by the man, the total strength of all the characteristics covered in the instrument. By itself, a single raw score is seldom of much value to the classification officer who has to use it, regardless of how accurate and dependable the test may be. See also paragraphs 65 through 69.

b. A raw score does not tell the degree to which a man possesses a given skill or aptitude *in comparison with other men in the Army* and is, therefore, no clear indication that he will do better or worse than others on assignment. A raw score does not tell what proportion of Army men stand higher or lower in regard to the trait under consideration. For each test, therefore, it is necessary to know the scores of other men with which a man is being compared and how they are distributed along the range from high to low. Classification officers have neither the time nor the facilities to collect this necessary data. Further, a time-saving and efficient technique for interpreting raw scores in terms of these data is required to make sound classification practicable in an Army of millions of men. The data concerning performances of other Army men are obtained by testing a standard reference population. The device for handy interpretation by the classification officer is the *Army standard score scale,* developed to show what raw scores mean in terms of comparisons among men (pars. 70 through 78).

34. Standard Reference Population

a. The population on which the experimental test is tried out (par. 32) is usually not the population on which the final form of the new test is standardized. This standardization sample must be representative of the population that will take the test. It is neither feasible nor efficient to give each new test to the whole population of Army men concerned. Sufficiently dependable information can be obtained by using a carefully selected sample. Each new test is given in its final form to a large sample of men selected to represent as accurately as possible the whole Army population for which the test was designed. Some of the difficulties involved in sampling have already been discussed in paragraph 30*a*. Further discussion of sampling is given in chapter 5. The size of the group varies considerably, depending upon the nature of the problem, the availability of groups, and the requirements of speed and economy. No practical advantage is gained by enlarging the sample at a high cost in time, energy, and personnel, since scientific control of its selection and the application of appropriate statistical techniques produce sufficiently dependable results. The representative group

to which the final form of the test is given is the standard reference population. The administration of the test to this population, and the statistical computations which follow, is known as standardization.

b. Sometimes it is possible to standardize an instrument on the entire operational population, rather than on a sample prior to adoption of the instrument for operating use. For example, no attempt is made to standardize the scoring of officer efficiency reports until all the reports for all officers for 1 year are available. This method eliminates the problem of adequate sampling since the total population is used. It also avoids the difficulties introduced by possible differences between experimental and operating conditions.

35. Minimum Qualifying Scores

The test score below which men may not be accepted for induction, duty assignment, or training course, is called a *minimum qualifying score.* Minimum qualifying scores for several different assignments may be set at appropriate points on the range of scores for a single test. The chances of success on a given job indicated by any obtained Army standard score can be computed. The minimum qualifying score is set at a point dictated by Army necessity. Thus, if it is desired that 80 percent of the men selected shall complete a course successfully, or perform satisfactorily in a given assignment, the minimum qualifying score could be set such that only men who stand a 4 to 1 chance of success, or better, will be selected. To select so high, however, may also mean that few will qualify. In establishing the minimum qualifying score, it is necessary to take into account the supply-demand ratio for the particular course or assignment in question. If the demand is small in relation to the supply, the minimum qualifying score can be set high and there will be low probability of failure among those selected. Where the demand is large relative to the supply it will be necessary to lower the minimum qualifying score in order to qualify more men. When this is done, a higher percentage of failures must be expected among those selected (ch. 5).

Section V. INSTRUMENT BATTERIES

36. The Need for More than One Instrument

The discussion thus far has been directed at the development of a single test. However, it is apparent that for many Army jobs the abilities required for successful performance are so complex that a single test, even a very good one, can hardly provide all the kinds of information needed to classify men. A battery of measurement instruments is needed to cover all the important abilities required.

37. Which Instrument Should Be Used

a. Before selecting or constructing measurement devices for a battery, a clear understanding is needed of the nature of the work for which selection is to be made. Consider a hypothetical example, the enlisted job of Combat Construction Foreman (E–6, E–7). The job duties might be summarized as follows: "Supervises combat construction, combat repair activities, and demolition activities, or acts as first sergeant of an engineer combat company." To be successful in this job, the man must have a good deal of information about the activities of a combat engineer company—mess, supply, transportation, records, Army regulations, intelligence reconnaissance—as well as information about the problems, materials, and procedures involved in construction, demolition, and camouflage. Much of this information can be tested by a written test devised for the purpose.

b. The Combat Construction Foreman also functions as an instructor of the men under him. He must actually demonstrate how to perform many of the tasks which are to be carried on. Tests of performance in installing and removing booby traps, using a mine detector, improvising charges for special demolition projects, and solving various kinds of rigging problems could be useful in determining how well he can do what he is supposed to do.

c. Not every man who is qualified to do the construction parts of the job can teach the job to some one else, nor have paper-and-pencil tests been developed as yet which give a useful

measure of effectiveness as a teacher or supervisor. It has therefore been common practice to try to obtain some measure of effectiveness in teaching and supervising by having competent observers make ratings of past performance in such duties.

d. In addition, the Combat Construction Foreman serves as a leader in combat situations. To get estimates of this important aspect, other kinds of measuring devices are needed, such as an instrument which provides information on his personal history, schooling, interests, hobbies, etc. (self-description blank). It would also be desirable to get some idea by direct observation of the man's typical effectiveness in dealing with people (standard interview).

e. At this point it can be seen that an adequate testing program for the job of Combat Construction Foreman would appear to require—

(1) A paper-and-pencil test of some length, covering a variety of kinds of knowledge.
(2) A performance test or work samples.
(3) Ratings on past performance.
(4) A self-description form or standard interview, or both.

Furthermore, each of these instruments must be valid for selecting men for this job; that is, the measures must be related to some independent measure or criterion of competence on this type of job. In addition, there should be a minimum of overlap in what each instrument measures.

f. The decisions reached thus far on what combinations of factors are important to success on a given job and how they should be measured have been based on judgment and logic. It is still necessary to determine whether these decisions are sound. Statistical techniques are available which will provide useful information on this point. A battery of a large number of instruments may give only slightly better prediction of job success than a reduced number of tests; the increased validity of the larger battery may not be of practical importance when weighed against the greater cost in time and convenience of using a longer battery.

g. This determination of the best combination of instruments for selection or classification of men is based on several considerations. Most important are the validity of each instrument and the degree to which each overlaps the others, that is, the correlations of each instrument with the other instruments (see ch. 5 for discussions of validity and of correlation). If two instruments are both valid but also extremely similar in what they measure, there is usually little point to using both in a battery. However, instruments which have some validity, but low correlation with each other, may make up a composite which is considerably more valid than any one of the instruments used separately. In such cases, a composite score may be obtained which represents performance on the battery as a whole rather than on each instrument separately.

h. Instruments in the battery may be given equal weight or differential weight. The weighting used is based on the results of statistical techniques. When the scores on the instruments are thus weighted, the composite score will yield the best prediction of success in job or training that can be obtained using the particular combination of tests making up the battery.

i. Selection of instruments for a battery and determination of the best weights are checked by administering the battery to a new group of men, representative of the group for which the battery was designed. This procedure, known as cross-validation, has been discussed with regard to test construction in paragraph 32.

j. There are other considerations involved in determining what tests should make up a battery. Practical problems such as the importance of the job, the number of men needed and available, time and expense, all are considered before a battery is installed for operating use.

38. Other Ways of Using a Number of Instruments

a. Instead of combining several instruments into a battery which yields a useful composite score, the instruments may be used separately in two ways—

(1) Minimum qualifying scores may be established for each instrument where a minimum level of competence is re-

quired for each of several aspects of a duty assignment or training course. This is not a very useful method since minimum qualifying scores are not fixed but should reflect variations in the supply-demand ratio. Furthermore, it is based on the assumption that a weak performance on one measure cannot be compensated for by very good performance on another.

(2) The instruments may also be used as successive screens. On the surface, this method is a very plausible one, but in practice it has serious weaknesses. One is that the order in which the screens are used can seriously alter their effectiveness. Another is that separate minimum qualifying scores are needed, and, as already emphasized, these are not fixed. Other difficulties are present, not the least of which is the temptation to add more screens than is profitable.

b. Considering all the advantages and disadvantages, it is usually better to use a battery of instruments as a composite than to use the instruments independently or as successive screens.

c. In general, the basic procedures for constructing a battery of instruments are the same as those for a single instrument, which might be conceived as a battery of items. For instance, just as the scoring of a single instrument is converted to a standard scale of measurement, so should the composite score be made meaningful.

39. Continuing Studies

When an instrument is released for field use, its usefulness remains a concern of the testmaker. Even if a test has been shown to have near perfect validity, it may still be necessary to check on it from time to time. The nature of a duty assignment or school course may be materially altered as a result of improvements in equipment and tactics. The nature of the Army population may change from a peacetime status to a mobilization status. Supply-demand ratios change. All these may affect the usefulness of a test. Follow-up studies need to be made from time to time to check the subsequent job performance of men who scored high and those who scored low on the test. Only through the study of accumulated evidence can the Army be sure that the test is doing what it is supposed to do, or ascertain when improved tests and procedures are necessary.

Section VI. SUMMARY

40. Steps in Instrument Construction

a. Study of the classification problem to determine the need for instruments, practical considerations influencing the structure of the instrument and the characteristics to be measured, and the specific way in which the scores will be used.

b. Design of the instrument in accordance with its purpose.

c. Construction of items.

d. Field trial of experimental form.

e. Analysis of results of field trial.

(1) Determination of validity, difficulty, and internal consistency of items.

(2) Preparation of scoring key.

f. Selection of items for final form.

(1) Use of indexes of validity, difficulty, internal consistency.

(2) Scoring based on selected items only.

g. Verification of scoring key.

(1) Application to a second group.

(2) Revision of items.

h. Computation of validity and reliability of revised instrument as a whole.

i. Preparation of operating form of test and procedures.

j. Establishment of scale of measurement.

(1) Standardization.

(2) Minimum qualifying scores as dependent upon supply-demand ratio.

k. Instrument batteries.

(1) Selection of instruments.

(2) Composite, independent, and successive screens.

l. Continuing verification studies when appropriate.

CHAPTER 3

CRITERIA

Section I. CRITERIA AND VALIDITY

41. Definition of Criteria

a. In chapter 2 it was shown how a newly prepared test is "tried out" to see how effective it is in differentiating between competent and incompetent men in a given occupation, or between men who are "good" on some characteristic and those who are "poor." But how is it decided which men shall be considered competent and which incompetent? What does it mean to be "good" or "poor" on some trait? It is clear that there has to be some standard to go by.

b. The measure of competence or of "goodness" used in making these distinctions is the *criterion measure.* To say that there is another measure by which the "test is tested" gives rise to a whole series of questions: How is this criterion decided upon? How can a measure of this criterion be obtained? How is it known whether or not it is a satisfactory measure? What is the essential difference between test and criterion? These and other problems will be discussed in detail in the following paragraphs.

42. Part the Criterion Plays in Validating a Test

a. The Criterion as Representing the Purpose of the Test. Determining the validity of Army personnel measurement instruments boils down to two basic problems, one somewhat theoretical, the other extremely practical. The theoretical problem is whether the test actually measures the trait or ability it was devised to measure. The chief limitation in the use of this approach is that even if it can be established that a certain trait is measured by the test in question, it still has to be established that the trait as measured is important to success on the job. The name of a trait or characteristic, in fact, means little. A man who is honest in money matters may not have the same scruples about taking advantage of his fellows when all are in line for promotion. The more practical question is: Is what the test measures, regardless of what it is called, important in performance of a given type of job? What must be ascertained is whether the test can provide a reasonably accurate means of estimating how good a soldier a man will be in the spot for which he is being tested. Only a measure of his performance on the job or in training can demonstrate whether or not the test is contributing as it was intended to do to the effectiveness of the selection process.

b. Criteria as a Check on Theory. The practical approach, too, has its limitations. Experience in research on the development of tests has shown that even after the most careful analysis of a job it may not be clear just what is required for success on that job. It is here that a criterion is most important. It is by checking theory about what is important to such success against a criterion of success that theory can be verified or errors in reasoning detected.

43. How Criteria Contribute to Classification

a. Each man could, it is true, be placed in a job or assigned to training in a specialty for which, by surface evidence, he seemed suited. If after 6 months or so he was doing satisfactory work or making satisfactory progress in training, his assignment could be made permanent. If not, he could be reclassified to some other occupation. That would amount to using

the criterion itself as a basis for classification. What is wrong with that?

b. In the first place, it would be a time-wasting process. A man who was misclassified would have lost 6 months or more of training time. More important, nothing would have been learned about the man's ability to do other jobs, nor would anything have been found out about other men who might perform much better in that particular job or training assignment. Even in the case of those who met the minimum performance standards for the job, it is not safe to assume that they have been properly classified. They might better have been assigned to some other occupation.

c. Thus, the use of tests instead of criteria not only saves time and expense but provides a means by which the whole classification procedure in the Army can be integrated. It is only by having these advance measures of probable success for each man that the process of matching manpower and Army needs can

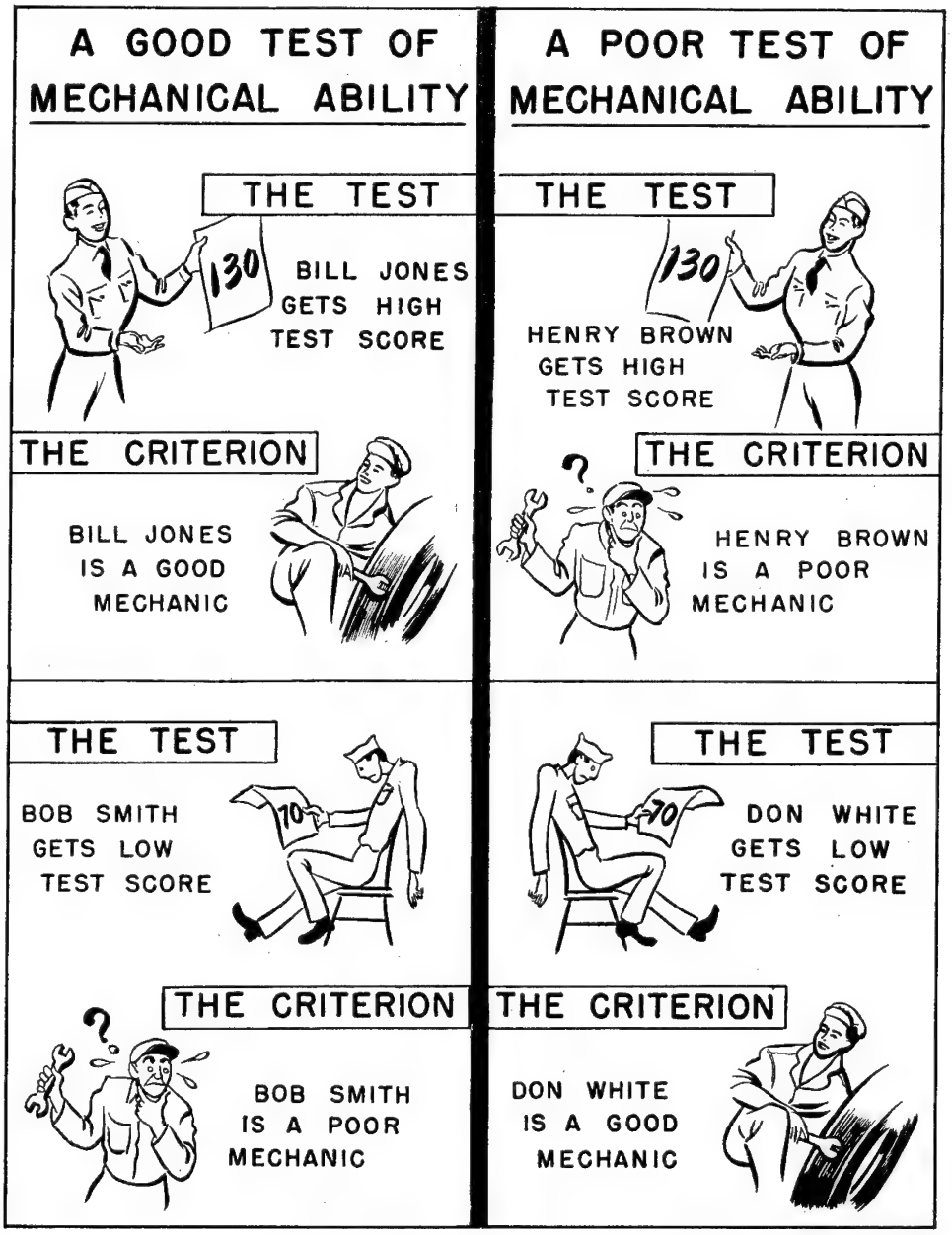


Figure 6. A criterion tests a test.

go on efficiently. On the other hand, these advance measures would be impossible to obtain without the use of criteria.

44. How the Criterion Measure Is Used

a. To illustrate the part the criterion plays in the development of a personnel measuring instrument, consider the problem of finding out how effectively a set of tests will pick men who will do well as auto mechanics after an appropriate training course. The experimental tests might include a mechanical aptitude test, a test of automotive information, a test of ability in visualizing how patterns fit together to form objects, and a test of how quickly a man can assemble some mechanical device such as a bicycle pump. These tests would be given to a group of men who had undergone training for the job of auto mechanic and were now assigned in that military occupational specialty.

b. At the same time, criterion measures would be obtained for each man in the group. Suitable criteria for this job might be performance ratings for each man by both his superiors and men working with him. How this is determined is discussed in paragraph 62. It could then be determined how closely scores on each of the tests correspond to the criterion scores, or, more accurately, whether the ranking of the men on the tests is the same as their ranking in ratings of performance on the job. This comparison of rankings is known as correlation, and the computation of the degree of correspondence between scores on a test and scores on a criterion gives a measure of the validity of the test (ch. 5). If the correspondence is close, that is, if men are in pretty much the same order on both measures, the test is said to be valid for this particular job. When the test, thus tried out against a criterion and found to be sufficiently valid, is given to groups of unselected men, there is considerable likelihood that the men who score high on the test will give good performance as auto mechanics after training, and the men scoring low on the test will perform poorly. The test is then said to be a good "predictor" of success as an auto mechanic. The degree of correspondence between the criterion and each test may be found in this way; or, as explained in chapter 2, the tests may be treated as a battery and a composite score on the tests, with each test weighted according to its validity and degree of relationship to the other tests, may be used.

45. Importance of the Criterion

It is necessary to determine at this point whether a certain test will be used for a particular purpose or which of several tests will be retained in a battery. Since this decision is made chiefly on the basis of the relative validity with which the tests predict the criterion, it is in effect the criterion which determines which tests will be used and which discarded. If the criterion measure is faulty, if it is not truly related to the essentials of job success, the wrong tests may be used. The men selected by these tests will have been selected on the wrong basis, and little will have been accomplished toward improving the method of selection for the job. The criterion, then, has a decisive effect upon the outcome of a selection procedure.

Section II. CHARACTERISTICS OF ADEQUATE CRITERIA

46. What Makes a Satisfactory Criterion

a. Since the criterion plays such a decisive role, it is essential that it be adequate and that it be adequately measured. If the criterion is to be adequate, it must be tied to the purpose of the job. It is necessary to determine what is required for success on the job and to determine how to measure the desired requirements for success in that job. For example, if it is desired that recruiters obtain as large a number of enlistments as possible regardless of the quality of the enlistees, the appropriate criterion measure is a production record. If it is desired that the recruiter enlist only high quality men, and not mere numbers, a production record is not satisfactory criterion measure —the criterion must emphasize the recruiter's ability to size up men. Similarly, if a rifleman is required to be primarily an expert in marksmanship, a production record—target scores—might be the appropriate criterion measure. If he is required to train recruits in handling a

weapon, ability to teach marksmanship would be the appropriate criterion measure.

b. Another aspect of the problem of determining whether a criterion is adequate is the relation of the job to the ultimate mission of the Army. Thus, the criterion of success of an infantryman in combat is performance in combat. If there is no combat, obviously no measures of the ultimate criterion can be obtained. It is necessary then to use success in garrison duty or training as the criterion.

c. A criterion of success on a job must be related to the purpose of the job. It is then necessary to determine what characteristics are important to accomplish the purpose of that job—characteristics which may *NOT* be important in other jobs. Thus, "military bearing" may be important in an infantryman on garrison duty. It may be of minor significance in an infantryman in combat.

d. It is apparent that considerable attention must be directed at studying possible criterion measures before it is decided which of them to accept (par. 62). Such studies are intended to discover which possible measures possess the necessary characteristics for use in criterion measures. These characteristics will be considered next in greater detail.

47. Comprehensiveness and Weighting

a. The criterion measures should cover in some way the important aspects of the performance, whether it be in a training course or on the job. For convenience, the discussion will refer only to job performance; it should be remembered that the discussion applies to training performance as well.

b. The important elements of the job must be analyzed by careful study. This can be done by interviewing men on the job and those who supervise them, by observing men at work on the job, or, on occasion, by actually performing the job long enough to discover its more subtle aspects. This job analysis must uncover not only the particular operations performed but also the personal relations involved. What is even more important than identifying the job elements is to determine their relative importance. There is little point to discovering that an infantryman must know how to read English and how to read military maps if it cannot be determined whether both skills are important for the infantryman and whether one is more important than the other. Both elements might be in the criterion but the more important skill would be weighted more.

c. On many jobs, more is involved in success than a breakdown into individual job elements will show. This is most apparent in jobs where the mission must be accomplished by working through others. The only criterion measure that will suffice is one which in some way takes account of how skill in dealing with other people influences all other elements of success in such a job.

d. Looked at from the point of view of validating a particular instrument, it may turn out that the instrument is closely related to one criterion measure but not to another. This condition would not necessarily mean that the instrument is of no value. It might well mean that the criterion measure which is not correlated with the predictor instrument is an inappropriate criterion.

48. Freedom from Bias

a. The criterion data should also be free from bias, that is, free from the influences of any factors, other than chance, that are not pertinent to on-the-job success. A man's score on the criterion may be influenced by factors which have little or nothing to do with how well he does his job. For example, he may have produced less work than another man, not because he cannot, or will not produce more, but because he is working with poorer equipment, or because he has to make up for someone else's mistakes, or because he gets less to do than the others.

b. Criterion bias often occurs when a rater knows how the man he is rating scored on the test that is being tried out. Knowing that the man scored high on the test, he may consciously or unconsciously rate him high on the criterion, regardless of his actual performance. Bias of this nature will make a test seem more valid than it really is.

c. Again, an officer who is rating his men for criterion purposes may know that a man makes a good appearance and talks well, and

therefore have the feeling that he does everything well. He may judge a man's skill in close order drill and infer from this that the man is doing well in squad tactics, instead of observing him in squad tactics and rating him on those duties. This type of bias is known as "halo," and will be discussed more extensively in paragraph 59*a*.

49. Consistency of the Criterion Measure

One general requirement of a criterion measure is high reliability. Criterion measures should be consistent from one time period to another. A rating obtained in October should not be too different from one obtained for the same man in May, if neither the job situation nor the man has changed drastically. Nor should there be too great a divergence between the criterion ratings when a man is rated by different raters within the same time period. The standards of consistency are considerably more relaxed for ratings than for test scores, but even here if there is not a reasonable consistency the measure is worthless as a criterion. Obviously, a characteristic which varies markedly cannot be measured consistently.

50. How Criteria are Selected

a. In any validation study, the criterion measure used is the one which is expected to prove most useful and practical. In making the selection, the type of instrument to be validated is considered and the purpose to which it is to be put after it has been tried out. Is it intended to screen out only the poorest men? Is it to select only a few best men for promotion? How specific is it to one particular type of job? Another consideration is the importance of this purpose and how much time and expense can be justified in obtaining—or developing—what is considered an adequate criterion measure. Timing, too, is important. How long can be spent in investigating the various possible criterion measures? In view of all these aspects of the problem, which of the available or obtainable measures related to job success will be the most satisfactory?

b. Some measure of job success may stand out as the most accessible and as bearing an obvious relationship to job success. To accept it as criterion without investigation, however, may lead to attempts to validate an instrument against a misleading standard. In developing tests to select driver trainees and Army drivers it was agreed that the criterion was safe driving. It would have been easy to accept one of two criterion measures which seemed clearly related to motor vehicle safety and, at the same time, fairly easy to obtain—accident records and road test scores. Both of these, however, had turned out to be open to objections. Accident records, the most obvious and easily obtainable measure, were dependent upon too many factors other than the safe driving practices of the motor vehicle operator. Road tests, another seemingly logical device, failed to sample a driver's behavior in potentially dangerous situations out from under the surveillance of the examiner. Both measures were unstable from period to period. In view of the shortcomings of these measures, a special study was set up to develop an adequate criterion measure in the form of ratings by the drivers' supervisors and associates. In developing the rating procedure special attention was given to narrowing the content of the rating to observable—that is, ratable—behavior important in safe driving. The result was a more reliable evaluation of safe driving against which driving proficiency or aptitude tests may, in the future, be validated. This illustration points out the importance of conducting criterion studies before selection of the appropriate measures is made (par. 62).

Section III. KINDS OF CRITERION MEASURES

51. General

a. Criterion measures usually belong to one of the two following types: an objective measure of job performance; and a subjective rating of success in an assignment by those in a position to judge. Which of these two types is selected depends, as will be seen below, not only on their availability, but also on the purpose to be accomplished.

b. The criterion measures termed objective include production and cost records, scores on tests of job knowledge, salary, or rank. In spite of their seeming advantage of objectivity, they have, in general, proved less satisfactory as criterion measures than have ratings. The various types of criterion measures will be considered in some detail.

52. Production and Cost Records

a. Examples of production records that may serve as criterion measures are quantity of work produced, quality of product, number of errors. Such records are attractive as criteria because of their relative objectivity and their very apparent relationship to on-the-job proficiency. Only a careful examination reveals that they may not be as objective as they seem. And while they usually do measure some element of job success, it is not always the most important one. Amount produced depends in large measure on opportunity to produce. Quality in workmanship depends, in part, upon the condition of the materials and tools with which the work is done. These are matters which the individual can usually do nothing about but which may bias his record.

b. It is sometimes possible to keep variations in assignment and working conditions under control or to allow for them in some way in obtaining the criterion measure. Such a measure then gives a partial criterion of success on the job, and it may be a significant part. With a typist, the most important consideration may be how fast and how accurately he can type. With more complicated and responsible positions, such a measure would represent at best only the most routine parts of the work. In practice, production records are commonly used in combination with other criterion measures.

c. In the Army, there are few jobs with which such a measure is appropriate, since there are few jobs where production is the main requirement. Production criteria have their main usefulness in situations where opportunity to produce may be fairly uniform.

d. Cost records may be considered a form of production record that is more useful than output records. In general, cost records should indicate how much time and money the organization must spend on a man for him to achieve a certain level of production. The usual production records do not indicate, for example, how much time a supervisor must spend in obtaining production from a subordinate, nor, to take another example, how much it costs to rectify errors made by a particular man. This type of criterion measure, the "dollar criterion," is a recent development and not much is known about it yet. Ratings on supply discipline may be considered a limited recognition of the importance of cost to the Army. Beyond this, the "dollar criterion" has not been used in the Army thus far. The principle involved, nevertheless, is an important one.

53. Status or Position

a. Sometimes status or position is considered for use as a criterion. A man who holds the grade of sergeant may be thought of as representing a higher level of competence than a corporal. Stated more generally, grade, echelon, and other indicators of official status may be considered as representing levels of responsibility and these, in turn, as representing levels of competence. A criterion measure of this sort might be obtained on a group of men of various grades whose test scores would be compared with their grades to see if the men with the high test scores are the men who hold the high grades.

b. If such a criterion measure is used in a follow-up study—that is, a study to determine whether the personnel instrument used to select the men is really predicting their relative success on the job—certain precautions are necessary. The predictor measures must have been taken before the men start in on their assignments. Results on the selection test should not be made known to those responsible for assignment and promotion, or their promotions may partially reflect this knowledge. In that case, status would be a poor criterion.

c. The difficulty with this type of criterion measure is its assumption that status is determined primarily by the competence of the individual. Sometimes this is true; sometimes it is not true. The whole question of equality of opportunity, of equal responsibility associated with equal rank, enters in. Before status or position is accepted as a criterion measure,

the basis on which it rests must be carefully examined.

54. Work Samples

Performance in a sample of normal work is sometimes used as a criterion measure. If the work is the sort that yields production records, the production records themselves may be used. If not, some other way of evaluating the work sample is necessary. This is likely to be a rating. In either case, the work sample performance is of doubtful value unless it is established that it does adequately represent the normal work. If not, the work sample performance, like production records, would have to be combined with other measures to provide a more comprehensive measure of effectiveness on the job.

55. Course Grades

Grades in Army service school courses may be appropriate criterion measures for instruments designed to pick men who are likely to succeed in acquiring certain knowledges and skills. It should be pointed out that such grades usually belong in the class of subjective measures. They are usually based in part on the instructor's judgment of how well the man has done in the course. A number of schools have adopted grading systems based on an objective test or a series of such tests given at various training stages. In these, greater confidence can be placed. Even when objectively arrived at, however, course grades in specific subjects are of doubtful value as criteria. They usually fail to take into account many elements of job success such as performance under unfavorable as well as favorable conditions, or the application of knowledge to an unfamiliar problem. On the other hand, it may be argued that while the ultimate objective is success in job performance, successful completion of the training course is a prerequisite to such success. School performance may then be thought of as another predictor of job success and, as such, to have validity as a criterion. The purpose of the school course in question and the basis on which its course grades are determined will usually give a clue as to the value of its grades as criterion measures.

56. Ratings

All the measures discussed above fall short in some way of what is desired in a criterion measure. At the present state of knowledge, the type of measure which most nearly approaches the desired scope for criterion use is the rating. Ratings are familiar as devices for evaluating men so that appropriate administrative actions may be taken. Administrative ratings or efficiency reports are discussed in chapter 10. Criterion ratings and administrative ratings have many similar characteristics. However, there is one great difference between them which arises from the fact that criterion ratings are confidential ratings which are not available for administrative use. It is the problems which arise in obtaining and using ratings as criterion measures which will be discussed in this chapter.

a. Ratings as the Best Available Criterion Measures. Ratings are used as criterion measures because they frequently are the most nearly adequate measures available for this purpose. The rater in appraising a man's performance will, in general, consider it in the light of the responsibilities and opportunities that are a part of the job. Unlike a production record, a rating does not disregard the conditions which can make each job somewhat different from every other job.

(1) A rating can be an over-all measure. Frequently that is the kind of criterion that is necessary. For example, personnel actions involving infantrymen are seldom concerned with competence in some particular element of the job. An infantryman is not considered competent just because he is a good marksman, or a good map reader, or good at rendering first aid. There are men who are good marksmen but who are, nevertheless, not good infantrymen. To obtain criterion measures of competence as infantrymen—an over-all measure—the only technique available at present is the rating technique.

(2) A rating involves the personal re-

actions of the rater to the ratee. Some of this is undesirable and represents a source of bias. Yet when job performance involves the need to work with and through other people, it is important to have a measure involving personal reactions. In spite of the difficulties encountered in attempts to get ratings that are both valid and reliable, ratings continue to be the criterion measure most frequently used.

b. Improvement of Criterion Ratings. The advantages of ratings as criterion measures are, in part, offset by their limitations. It is, therefore, important to discover means of reducing these defects. Considerable research effort is directed at improving ratings so that their effectiveness as criterion measures may be increased (par. 61).

Section IV. RATINGS AS CRITERION MEASURES

57. Ratings are Recorded Judgments

The term "rating" is used in a general sense to mean all methods of expressing judgments on the adequacy of a performance. All ratings are essentially recorded judgments. Perhaps the oldest method of recording a judgment is the essay, in which the rater expresses his judgment in his own words (ch. 10). The familiar rating scale may be considered as an attempt to improve ratings by greater systematization of the method of recording the judgments.

58. Description of Rating Scales

a. There are many forms of rating scales. In its simplest form, the scale is a series of steps representing levels of performance or degrees of a characteristic possessed by the ratee. Descriptive terms may be supplied and numerical values assigned to each step of the scale. Each person rated is then placed at the level where, in the judgment of the rater, he properly belongs.

b. Such a scale is most often accompanied by a scale of measurement. The graphic rating scale is probably the most serviceable means of obtaining a rating. Indeed, the "cross-on-a-line" method has been so widely adopted that the term, "graphic rating scale" has been generalized to include all rating scales (except checklists) whether or not they are truly graphic. Such a scale consists of a line divided into intervals representing varying amounts of the characteristic on which the rating is made. Descriptive phrases may be placed along the line from one extreme to the other to define the points of the scale. A variation in this procedure is the man-to-man rating scale. Raters are asked to choose specific persons of their acquaintance as typical of each of the scale positions. The rater thus has a standard—set by himself, it is true—and can place the ratees by comparing them with the persons he has selected to represent the various points on the scale.

c. There may be one scale or many. With several scales it must be determined how much each shall count in the composite rating. The final numerical score may be converted to a standard score by the method explained in chapter 4.

59. Limitations of Rating Scales

Rating scales have become the traditional method of rating. Nevertheless, they possess certain serious defects which cannot be overcome simply by preparing a new form. The basis for these defects lies less in the form than in the rater. In the study of the rater, four general human frailties affecting ratings have been well established—tendency to rate on general impression (halo); tendency to keep all ratings close together on the scale (lack of dispersion); tendency toward leniency (rating high); and differences in raters' standards (differences in frame of reference). The combined effects of these tendencies for the traditional type of scale is to bring into operation quickly the law of diminishing returns in attempts to lengthen rating scales; to reduce the spread of rating scores, often to the point of making them useless for any practical purpose; and to obtain scores which are often more a reflection of differences in raters than in ratees.

Type 1

In this section you are to make a series of over-all evaluations. Rate the man on each of the qualifications below. Mark an X in the appropriate space on each of the scales to show your evaluation of the man. Judge each man by comparing him with all the men you know of the same grade and having the same responsibilities

1. ADJUSTMENT: Degree to which he is able to meet situations without prejudice and without emotional upset.

POOREST — BEST
1 2 3 4 5 6 7 8 9 10

2. COOR[illegible]egree to which he is able and [illegible]ith others.

POOREST — BEST
1 2 3 4 5 6 7 8 9 10

POOREST — BEST
1 2 3 4 5 6 7 8 9 10

Type 2

Instructions to rater: Consider carefully each of the five descriptive paragraphs below; then, on the basis of your observation of the trainee whose name is entered above, decide which of these paragraphs best describes his work on the project and place a check-mark (✓) in front of that paragraph.

☐ (1) Could not complete job even with major assistance from instructor. Did not know the relative parts of his job either by definition or use. Had no understanding of why the job was be done.

☐ (2) Was able with difficulty to complete parts of the j[illegible] Had an idea what to do but lacked sufficient ins[illegible] dexterity to complete all parts of the job. little of why he did the job.

☐ (3) Had a general idea of what was to b[illegible] but with minor errors of omissio[illegible] starts, changes, and repetiti[illegible] his product.

☐ (4) Completed th[illegible]

Type 3

Military Bearing and Neatness: Slovenly / Not Neat — Poor Posture / Physical Defects — Clean; Orderly / Average — Very Neat / Well-Groomed — Prepossessing / Attracts Attention

Physical Adaptability: Abundant Energy / Athletic — Well-Coordinated / Vigorous — Average / Does Not Attract Attention — Exerts Himself with Effort; Listless / Awkward and Clumsy — Lazy; Stands on Sidelines

Group Adaptability: Repellent / Is Avoided by Others — Colorless / Does Not Attract — Accepted / By Group — Congenial / Is Sought After — Interested But / Requires Supervision — Not Interested / Resistant

Attitude: Dynamic; Self-Starter — Determined / Resourceful / Enthusiastic — Interested / Shows Desire to Succeed — Has Difficulty In / Expressing Himself; Uses / Bad Grammar & Language — Gives / Directions Well — Takes The / Lead Sometimes — Clear / Direct — Depends / On Others — Convincing / Forceful / Persuasive — Never Wins Group / Support; Creates / Antagonism — Well-Balanced / Stands Fast — Meets Emergencies / Calmly

Figure 7. Examples of three types of rating scales.

These problems are all present in obtaining ratings for criterion purposes.

a. Tendency to Rate on a General Impression —Halo. The rater's tendency to rate a man at or near the same level on all characteristics is a strong factor making it difficult to obtain useful ratings. The rater may have noted some feature of a man's behavior—an air of general efficiency or a forceful speech at a meeting. The generally favorable impression he has already formed may dominate his judgment when it comes to rating the man on his ability to organize a job effectively, or to lead a scouting patrol in a combat zone. Yet the behavior on which the impression was formed is not necessarily related to either of these abilities. It is quite possible for a man to appear alert and efficient and to talk well, and at the same time lack the ability to get work done or to be effective in combat.

(1) Attempts to reduce the halo effect have not been very effective. A practical adjustment can be made by limiting the number of different characteristics to be rated. Recent studies have shown that not more than five aspects of human behavior can be rated at the same time with any degree of independence.

(2) Certain other steps have been taken to reduce the effects of halo. Careful definition of the behavior to be rated, appropriate directions to the rater, an interval of time interposed between ratings on the separate characteristics, and education of the rater may promote a more careful and independent consideration of the separate scales and reduce somewhat the intercorrelations among the measures. However, the value of such lowered intercorrelations must be judged in terms of the purpose of the rating. In fact, if the purpose of the rating is considered, the rating may be reduced to a measure of a simple factor—for example, promotability—regardless of the number of scales used. This is true for criterion ratings as well as for administrative ratings.

b. Tendency Towards Leniency. Raters tend to give ratees the benefit of the doubt and to rate them higher than their performance warrants. This tendency, although operating in criterion ratings, is not usually so pronounced as in administrative ratings. Two facts arising from the nature of criterion ratings aid in reducing leniency in criterion ratings. One is that raters know that the criterion ratings are not available for use in administrative actions affecting the ratee. The other is that the rating instructions point out that the accuracy of the criterion ratings is critical in determining the nature of the personnel instrument which may be applied to them and their associates, and that the rendering of criterion ratings with a minimum of leniency is the raters' most effective way of contributing to the development of the instrument.

c. Tendency to Concentrate Ratings—Lack of Dispersion. No matter how many points are laid out on a rating scale, raters generally fail to use them all. Raters tend to rate men alike instead of differentiating among them. Except in rare instances no two men are of the same value. Sometimes the differences are relatively small, but careful observation will usually show the existence of differences. The amount of differentiation which the criterion ratings show has a direct bearing on the validity of the instrument being tested against that criterion. Thus, if the criterion ratings are held to a narrow range and this restriction is due to a bias of the raters rather than to any real lack of differences among the ratees, the results of a validation study may be misleading. The instrument may appear to be unable to differentiate between effective men and ineffective men; yet if the criterion ratings were not so concentrated, the instrument might well demonstrate its differentiating power. Generally speaking, criterion ratings show somewhat more dispersion than do administrative ratings, a result of the raters' understanding of the nature of criterion ratings.

d. Differences in Rater Standards (Differing Frames of Reference). Many ratings obtained are as much a product of differences in raters as of differences in ratees. Different raters apply different principles in rating. Each rater has his personal standard of what is good or

poor performance. He has likes and dislikes which influence his opinions of others. Because this type of rater bias may stem from such a variety of causes, it has been particularly hard to deal with. No attempts at erasing these differences have greatly succeeded. Rater agreement may be substantial. It is never complete.

60. Other Methods—Ranking, Paired Comparison, Nomination, Rank Comparison

In addition to essays and rating scales, other methods may be used as criterion measures (*a* through *d* below); their advantages and disadvantages are discussed in *e* below. The use of these methods in administrative ratings is outlined in chapter 10.

a. Ranking—Order of Merit. Ranking consists of placing men in order of merit. Each man is assigned a number, the one standing highest in the group is assigned number 1, the next, number 2, and so on through the group. The last number assigned is equal to the size of the group. When a number of raters evaluate each man, as is usual in criterion rating, an average is computed for all the ranks assigned the ratee by all his raters. In obtaining criterion measures, efforts are usually made to ask all raters to rank approximately the same number of ratees, thus avoiding having one rater rank only 2 or 3 men and another 15 or 20 men. The rankings may be converted to rank in 100 or rank in 1,000, or to a standard score. The standard score is the more useful conversion for criterion purposes.

AN EXAMPLE OF RANKING AS A RATING METHOD

The following named Privates in your unit are eligible for promotion. You are to rank these men in the order in which you think they merit promotion. Indicate the person you think should be first promoted with a "1", the next preference with a "2", and so on through the entire list. This list is submitted to you in alphabetical order.

NAME	RANK each person here
Brewer, James	7
Byer, Henry	4
Cantor, Alfred	3
Klein, David	1
Main, Christopher	5
Porter, Patrick	6
Wiltz, Ralph	2

J. J. Jones
Signed
Sgt., Co. C

Figure 8. An example of ranking as a rating method.

b. Paired Comparison—Comparing Two Men at a Time. The rater compares two men at a time and decides which of the two is the better. Each man is compared with every other man and the number of times each man is judged the better of the pair constitutes his rating. With other than small groups, this method is too laborious for practical purposes.

c. Nomination—A Variant of Ranking. Instead of ranking all men in order of merit, criterion raters are sometimes asked to name a certain number as best and an equal number as poorest. The score is the number of times the ratee was nominated as best and as poorest. A similar method, the "sociometric method," requires each rater to name the man he would want most for a particular duty and the men he would want least.

d. Rank Comparison—Ranking and Paired Comparison. To deal with relatively large groups, a method combining the order-of-merit method and the paired comparison method may be used. The group is divided into smaller groups of 15 to 20 men, and the men in each group are ranked. The two sets of rankings are merged by a modified paired comparison method and provide one order of merit.

e. Advantages and Disadvantages. The methods described in this paragraph reduce all the limitations described in paragraph 59 except one. The criterion scores obtained with these methods are not concentrated as are scores obtained with the traditional types of rating scales, since ratees are compared with other members of the group and some type of order of merit is established. As has been pointed out, such spread is highly important in a criterion measure. It directly affects the size of the correlation between the criterion

rankings and the instrument being validated. The spread obtained with rankings, however, may give a false impression that the men all differ from one another by equal amounts of the characteristic in question. This impression should not be allowed to obscure the fact that some of the discriminations apparent in rankings may have little or no real basis. It is comparatively easy to rank the four to five outstandingly good or poor individuals in a group. It is more difficult to assign accurate rankings to those in between. The rater actually may know of little or no real difference among them. The resulting rank order may be an artificial one. The raters' tendency to leniency is overcome by the required spreading since some ratees must be ranked as best and some as poorest. Similarly, differences in raters' standards are reduced. Of course, a particular rater can be lenient in evaluating a particular ratee or his standards may shift as he evaluates a particular ratee, but all the raters cannot be lenient with all their ratees nor can each rater use a different standard for evaluating each ratee. The one limitation which is not overcome by these methods any more than it is overcome by the rating scales is halo, the rating by general impression.

f. Conversion to Comparable Scores. The scores obtained by the methods described in *a* through *d* above usually cannot be employed directly as criterion measures. All raters do not usually rank the same number of ratees, thus affecting the numbers assigned to the lower ranks. It is therefore necessary to convert these scores to rank in a larger group or, preferably, to standard scores. Another difficulty is that a particular ratee ranked as, for example, the best by one rater may not be equal in quality with another ratee ranked as best by another rater. The only solution to this difficulty is the knowledge that the men are approximately comparable. Ordinarily, when it is known that ratees in one group are of a markedly different caliber from those in another group, ranking methods are not used. Special studies may be made to determine how the relative rankings compare with some absolute standard. As stated immediately above, conversion of scores, usually to standard scores, is necessary. It should be pointed out here that such conversions when applied to administrative ratings introduce the special problem of acceptability of the ratings to the rater (ch. 10). However, this problem of acceptability is not so important in criterion ratings since they are not used for personnel actions involving the ratees.

61. Improving Criterion Ratings

The limitations described in paragraph 59 have one net effect; they introduce systematic error into the measurement. The errors due to chance may be accounted for by appropriate statistical techniques (ch. 5). The systematic errors require special methods. Such errors are said to bias the ratings. To reduce bias in criterion ratings several methods are available.

a. Averaging. When a number of measurements of an object or a person are available, it is usual to average them. The average provides one convenient number to represent a variety of numbers. In addition to being a convenience, the average has another valuable property as applied to ratings—the biases resulting from the peculiarities of the individual rater can be reduced and a more stable measure obtained.

(1) In addition to partial canceling out of the effect of individual rater bias, the use of an average of a number of ratings improves the basis of the rating by combining what a number of raters know about the ratee. A better estimate may thus be obtained of his effectiveness. Of course, increasing the number of raters will be of no purpose if they are not competent raters of the ratee.

(2) It should be noted that if the individual ratings are concentrated within a narrow range, averaging will only accentuate the lack of spread and may lower the usefulness of the ratings for criterion purposes.

b. Guided Rating. Attempts to reduce bias in ratings may go so far as to provide actual guidance in making the ratings. An expert in rating methods works directly with the rater, or with groups of raters. Together, they reach an understanding of what qualities are to be rated and on what basis judgments are to be

formed. It is the role of the expert to help the rater think back over his observations of the man he is rating and weigh them as indicators of typical action or performance. His guidance is intended to help the raters avoid as much as possible the weaknesses which usually detract from the value of ratings. This method offers a practical means of obtaining improved ratings for criterion purposes. It is usually too time-consuming and expensive for adoption in obtaining ratings for administrative use.

c. Education of Raters. The human tendencies which operate to introduce bias into ratings are so ingrained that they cannot be overcome without serious effort. Attempts have been made to educate raters so that they will observe more carefully and rate more objectively. They have been informed that men differ in their abilities and that if their abilities were plotted on a graph, a normal curve (ch. 4) would result, if there were no bias in the measurements. When asked how they would distribute their ratings, raters said they would spread them out. Yet these same raters showed the characteristic pile-up of the higher ratings. Education in the form of knowledge of principles is not very effective in reducing bias of various sorts and more practical education is, at present, too time-consuming to be useful.

d. Special Instructions. In obtaining criterion ratings, special instructions are generally supplied to the raters, emphasizing the basic principles of accurate rating. Usually, these instructions are given orally by a technician who also attempts to clarify the instructions, if necessary, and to remind raters of the thoughtfulness and care required. This practice is similar to guided ratings except that it is applied to groups instead of to individuals. In general, such instructions have proved helpful but by no means do they eliminate the various sources of bias in criterion ratings.

e. Selection of Raters. Since the major source of bias is in the rater rather than in the particular rating procedure employed, it is natural that attempts be made to use as criterion raters only those who have met certain minimum requirements. These requirements are based on the belief that a rater, to be competent, must know what is required of the ratee and how well the ratee has met these requirements. Thus, it may be specified that the criterion rater must have a certain minimum length of service and a certain minimum length of work contact with the ratee. Command relationships are usually ignored and any rater who knows the job and the ratee's performance on the job is eligible. Superiors, equals, and subordinates may be used, if they meet the minimum requirements.

f. Rater Characteristics. Recently, studies have been undertaken to determine what rater characteristics are associated with ability to rate. In one study it was found that hard raters and easy raters rendered equally valid ratings—the raters placed their men at different parts of the scale, but they placed the ratees in much the same order. Another study indicated that ratings by enlisted men with Aptitude Area I scores (based on Reading and Vocabulary, Arithmetic Reasoning, and Pattern Analysis Tests) below 90 were not as valid as ratings by men with higher scores, even though special pains were taken with the low scoring men. Such studies, if continued, may disclose which characteristics are usable in selecting raters to give less biased criterion ratings. The possible use of such characteristics in selecting raters to give administrative ratings is limited and is discussed in chapter 10.

g. The Forced Choice Method. The forced choice method, described in chapter 10, may be an effective means of reducing rater bias. Like rating scales, the forced choice method yields a score which is not dependent on the size of the group. In addition, since the rater cannot tell how the ratee is going to stand in comparison with other ratees, the tendencies to be lenient and to concentrate ratings may be reduced. Although used in administrative ratings in the past, the forced choice method has only recently been used to obtain criterion ratings. As yet, no information is available on its effectiveness in criterion ratings.

62. Criterion Studies

a. General. The critical importance of criteria in determining the usefulness of personnel measuring instruments and their effect on the management of military personnel makes it essential that the measures used as criteria

be properly selected. This point has been emphasized in this chapter several times. To select the proper measure is not always easy, and considerable research effort may be needed.

(1) Sometimes this effort is directed at discovering whether existing measures are suitable for use as criteria. Too often, an existing measure seems suitable only until a more critical study is made of it. For instance, ratings by associates on over-all value may be available and would seem usable as criteria. However, investigation may reveal that these ratings are used for administrative purposes, that they are not confidential, that raters did not know enough about some of the ratees to rate them accurately, that raters made little effort to spread their ratings, and so on. What appeared to be a suitable criterion measure may prove not to be so. Consider another example—Leadership ratings may be available at a school and would appear at first glance to be suitable for criterion purposes. However, suppose it turns out that the raters were classroom instructors and that leadership ratings were highly correlated with academic grades. As criteria for validating academic aptitude tests, the leadership ratings might be suitable, but then the question would arise, are the ratings leadership ratings? As criteria for validating instruments intended to measure leadership potential, the leadership ratings could not be used unless it were known that academic performance and leadership were closely tied together, which is frequently not the case. In brief, then, considerable research may be required to determine if existing measures are appropriate for use as criteria, and, if not, to develop suitable criterion measures.

(2) Criterion measures may themselves be used to predict performance later on. For example, research may indicate that success in a particular service school is correlated with later success on the job. The measure used to evaluate success in the school may be the one used to determine the validity of the instruments used to select applicants for the school.

(3) Criterion research may yield an important by-product. The criterion measures may have been developed to determine the validity of certain personnel measuring instruments. They may also be used as yardsticks in other areas of personnel management. For example, they may be used to evaluate training outcomes, the effect of assignment policies on performance of the men, the effect of morale conditions on performance, and so on. Some modifications may be necessary but the basic content of the criterion measure could be used.

b. Relation of Criterion Measures to Purpose. It has been pointed out in paragraphs 42 and 47 that the criterion measure selected must be related to the purpose of the instrument to be validated. Sometimes this purpose can be established as a result of the knowledge of the job requirements and advice of job experts. For example, it can be readily determined that an artilleryman should know basic arithmetic and a test can be constructed to measure knowledge of basic arithmetic. The criterion measure could consist of evaluations of the ability to compute fire direction problems to determine how well the arithmetic test measures ability to handle fire direction arithmetic. But suppose, to follow the same example, it is desired to use the arithmetic test not simply to indicate how well men can handle fire direction arithmetic but rather how well men can perform as all-round combat soldiers. Another kind of criterion measure is called for, one which is more comprehensive. Further study may be required to determine the scope of the criterion measure and the method to be employed. Sometimes the use of a criterion is determined by policy. For example, instruments used to select officers should be validated against over-all criterion measures because of the policy of rotation. If ability to perform a technical speciality is required, the criterion

of over-all officer performance may be inappropriate. In general, then, both the purpose of the test and the purpose of the job must be understood in order to select the appropriate criterion measures.

c. Relations Among Criterion Measures. Sometimes a large number of possible criterion measures are available and it is desired to study the relationships among them as an aid in determining how they should be used. Measures which are highly correlated with each other may be examined to see which may be discarded to avoid unnecessary duplications. Those which are not correlated with each other are examined for suitability. They may be uncorrelated with other measures because they are poor measures, in which case they are dropped from further consideration. However, they may be uncorrelated with the others because they cover an aspect of performance not covered by the other measures. If this aspect is considered important in the successful accomplishment of the job and the measure is sufficiently reliable, it would be included as a criterion measure. As a matter of fact, considerable effort is directed at studying the correlations among the available measures to discover the genuine and important aspects of job performance. It is conceivable and has actually happened that none of the measures available for criterion use show any correlation with each other, in spite of the fact that the job requirements indicate that certain job aspects are not independent of each other. Put another way, measures which should be correlated with each other turn out not to be correlated. Such a finding suggests that the grading and evaluation systems employed are defective and points also to the possibility that the job duties or training content are not effectively organized. Until the necessary corrective steps are taken, there is little point in developing expensive measuring instruments to predict job performance when there is nothing to predict.

(1) There is another type of relationship among criterion measures which is of considerable importance—the relations among measures at successive stages in the Army career. Are the men who are good in basic training good in garrison duty? Are they good in service school? Are men who are good in service school good in combat? Many such questions may be asked—the difficulty lies in answering them. One of the difficulties arises from the fact that ideally a sizable group of men would have to be earmarked so that comprehensive personnel data would be available for the various phases of Army careers. A step has been taken in this direction with the introduction of an Army enlisted student evaluation report and an enlisted on-the-job data sheet to provide information on performance at service schools and performance on the job at a later date. One of the problems involved in such studies concerns the particular technique of measurement employed. Suppose, for example, that academic grades at a service school show substantial agreement with ratings of performance on the job for which the schooling is intended. This agreement would be especially significant since different measurement techniques are used. If, on the other hand, the academic grades show no agreement with the later ratings of job performance, no final interpretation can be made until it is known that the differences in measurement technique are not responsible for the lack of agreement. To obtain this information, studies may have to be conducted to determine the adequacy of the measurement techniques.

(2) One final point should be made regarding the relations among criterion measures. In view of the Army's mission to fight and win wars, it might be thought that the criterion measure that should be used is performance in combat. This point is recognized and studies to validate

various instruments against performance in combat are made when opportunity offers. However, it does not necessarily follow that the ultimate criterion—combat performance—is the only one that should be used. For one thing, combat criteria are obviously not always available, and during peacetime, when the Army still has a job to do, other criteria must be used. For another, combat criteria are not especially appropriate for all military occupations. For example, the closest to combat that automotive engine rebuilders may get is rear echelons and there is little point in attempting to get measures of performance under fire. On the other hand, there is little doubt that for riflemen a criterion measure based on performance in combat is appropriate. This discussion is another illustration of the principle that the selection of criterion measures must be related to the purpose of the instrument to be validated and the purpose of the job.

d. Criteria of Unit Effectiveness. The need to evaluate the effectiveness of small military units or teams (squads, platoons, companies) has long been recognized in the Army. More or less standard field problems have been used to aid such evaluations in connection with training and determination of combat readiness. However, the usefulness of such measures is not limited to this purpose, and because of their potential widespread value, studies have been undertaken recently to improve them, using scientific personnel measurement techniques. A test of this sort designed to measure the effectiveness of an armored infantry scout squad has been developed and is to be incorporated in an armored field manual. A similar test for infantry squads is under development. These instruments are based on field problems, but they differ from traditional field problems in that their design, development, and scoring have involved measurement principles discussed in this manual. They are just as useful as the traditional field problems for training exercises and as aids to administrative and operational decisions. They now also meet the demand for adequate criterion measures. Against these measures, personnel instruments and training and management policies can be validated in a somewhat different way than if they were validated against measures of individual performance. With such criterion measures of unit effectiveness, it becomes possible to study systematically the impact on unit effectiveness of varying the characteristics and training of its individual members. For example, should all the men in an infantry squad have high Aptitude Area I scores or will the squad be equally effective if a certain number have high scores but the others have fairly low scores? Is it necessary for maximum effectiveness of a unit that all the men attend service school or will the unit be equally effective if only certain men attend service school? Also, if a unit measure is available, the relationship of various alleged morale or motivational factors to unit effectiveness can be better determined. In short, recognition of the value of an adequate measure of unit effectiveness makes this an area of increasing importance in military personnel management.

63. Relation to Policy

The extended discussion of criterion problems should not be lost sight of in establishing policy regarding the use of personnel measuring instruments. The criterion measure plays a critical part in the development of an instrument, and should be kept in mind when establishing policy as an aid in guarding against overestimating or underestimating the effectiveness of the instruments. The criterion is a measure of purpose and hence determines the effectiveness of an instrument for that purpose. For other purposes, other criteria—and consequently other instruments—may be needed. In other words, the criterion should not be lost sight of in establishing policy to determine the use of certain measuring instruments.

Section V. SUMMARY

64. The Criterion Plays a Critical Role in the Development of Personnel Measuring Instruments

a. Criteria play a decisive part in determining the effectiveness of personnel measuring instruments. Criterion measures for validating instruments may also be used in other areas of personnel management.

b. The adequacy of the criterion measure is determined by the following:

(1) Relevance to purpose of instrument and job.
(2) Comprehensiveness and weighting.
(3) Freedom from bias.
(4) Consistency.

c. The kind of criterion measure most generally useful is a rating. The principal methods are—

(1) Rating scale.
(2) Paired comparison.
(3) Nomination.
(4) Rank comparison.

d. Each of the rating methods has advantages and disadvantages which must be understood. These are not necessarily the same as in ratings used for administrative purposes.

e. A major research interest is the development of methods to improve criterion ratings. The use of an average of a number of ratings instead of single ratings is a simple and effective way of increasing the value of ratings.

f. Criterion research is directed at studying available measures and developing new ones. In addition to criterion measures based on individual performance, attempts are being made to develop measures based on unit performance.

g. The nature of the criterion should be included in the factors considered when establishing policy governing the use of personnel instruments.

CHAPTER 4
THE MEANING OF SCORES

Section I. THE NATURE OF PERSONNEL MEASUREMENT

65. Measurement is Approximate

a. Measurement in personnel research is approximate rather than exact. However, it is not fundamentally different in this respect from measurement in other fields. The difference is one of degree, not of kind. An engineer thinks in terms of "tolerances" and specifies what tolerances he desires. He knows that when he says a metal bar is 33 inches long, he does not mean that it is exactly 33 inches long, but perhaps .014 or .045 inches longer or shorter than the stated length. The amount of difference which can be tolerated depends upon such factors as the additional time and expense of measuring the bar more exactly and whether the more exact measurement will make any difference in the use of the bar.

b. A similar situation exists in personnel measurement. A test score of 119 does not mean exactly 119. A score of 119 may mean, say, any value from 111 to 127, and the personnel psychologist may say something to the effect that in a particular case the chances are 40 out of 100 that the score is actually between 118 and 120, but 97 out of 100 that it lies between 111 and 127. The range of scores from 111 to 127 may represent an area in which we are almost completely certain that the man's true score lies. Factors affecting the size of these "tolerances" in personnel measurement will be discussed in this chapter. See also chapter 5.

66. Scores are Relative

It will be recalled (par. 7) that a number by itself is not a meaningful score and that standardization and validation are needed to provide the necessary meaning. A score must express a relationship between one man and the other men, either on a particular test, or on a criterion of value on the job. That is, the score must be interpretable as a "high," "low," "slightly above average," "average," etc. score on a test, or it must be interpretable as a score made by a "good soldier," "poor soldier," and so on. Scores, then, are relative.

Section II. TYPES OF SCORES AND STANDARDS

67. Adjectival Measures

a. A superior observes his subordinates and rates them as *excellent, superior, satisfactory,* or *unsatisfactory.* Another rates his subordinates as *outstanding, excellent, satisfactory, unsatisfactory,* or *very unsatisfactory.* This is a simple way of evaluating men. But is it as simple as it appears? What do the adjectives mean? Obviously, they do not mean the same things to all men—in the above example, *excellent* means "best" to one man, "next to best" to the other. Even if it were agreed to use *excellent* to mean only one thing, say "best," the problem is by no means solved. One man has high standards and very few of his men are evaluated as *excellent.* Another man has lower standards and a large number of his men are evaluated as *excellent.*

b. Even if differences in standards could be eliminated, there is still a serious problem. Suppose 100 men are rated on their value for a certain type of job and 15 are rated excellent. Suppose, further, that it was desired to pick the best 10 for a special assignment. Which of

the 15 "excellent" men are the 10 best? It is this question which illustrates a final difficulty in the use of adjectival measures for instruments involved in personnel actions—they indicate only very roughly how one man stands in relation to other men.

c. To sum up, then, adjectival *measures* are crude measures and are of limited usefulness in Army personnel actions.

68. Raw Numerical Scores

a. An improvement over the adjectival measure is the raw numerical score which may be a simple count of the number of correct responses made by an individual. The test answers are scored with a fixed key so that what the scorer thinks about the test or the answers is removed from the picture. Any two people scoring a paper should, except for errors, come up with the same total score. However, interpreting such scores is still quite a problem.

b. The difficulty of interpreting raw scores can be seen from the example below. Jim Brown is given three tests on which he makes the following scores:

Test	*Number of items*	*Incorrect answers*	*Correct answers*
Clerical speed test	225	50	175
Shop mechanics test	40	1	39
Vocabulary test	53	5	48

Brown missed only one question on the shop mechanics tests, but five on the vocabulary test and 50 on the clerical speed test. Yet his score of 39 in shop mechanics was lower than his score in the other two tests. How can the classification specialist determine from these scores the test on which Brown did the best? He can do so only by taking into consideration the difficulty of the test. It might be more of an achievement to get all but five questions right on a difficult test than all but one right on an easy test of the same length. Before Brown's three scores can be compared, they must be changed to some common scale to take account of different numbers of items, differences in difficulty of the tests, and differences in ability of the people taking the tests.

c. Similarly, it is difficult to compare two men on the same test on the basis of raw scores. Suppose, for example, it is desired to compare Brown with Smith who answered, say, 29 correctly on the shop mechanics test. Compared with Brown's 39, Smith's 29 is ten points lower. Is a difference of ten points a large one or a small one? Suppose, further, that in the clerical speed test Smith answered 165 correctly as compared with Brown's 175. Again, there is a difference of ten points, and again the question is asked, is this difference a large one or a small one? The only way to answer this question is, as before, to convert all scores to a common scale.

d. Raw numerical scores, then, may have the advantage of objective and uniform scoring, but they permit only a very rough comparison between men on one test or between one man's performances on two different tests. Types of standards that have been developed and are being used will be discussed in the rest of this section. It should be kept in mind, of course, that the computation of raw scores is always preliminary to the calculation of any more refined scores.

69. Percentile Scores

a. General. One type of standard is the percentile score by which one person's performance on a test is evaluated in terms of the scores of all the other people who took the test. Percentile scores may range from 0 to 100, when rounded to the nearest whole number. In certain instances, percentile scores below 1 and over 99 are presented with one or more decimals; use of the actual limiting values of 0 or 100 may thus be avoided. If a man gets a percentile score of 90, this means that his raw score was higher than the raw scores of 90 percent of the men who took the test, and equal to or lower than the raw scores of 10 percent of the men. If he receives a percentile score of 50, he excelled 50 percent of the group.

b. Advantages. Percentile scores offer a realistic way of interpreting tested performance because they compare the individual with the group. Since it is virtually impossible to determine the absolute amount of ability possessed by an individual, the determination of his relative ability is the most meaningful approach.

c. Disadvantages. Percentile scores are generally favored because they are easily understood. But a percentile system suffers from one

serious disadvantage. While it reveals what proportion of the total group tested an individual excels, it does not indicate *how much* better or poorer he is than any one of the others. This is true because percentile units are not equal to one another in terms of raw scores. For example, 5 percentile points difference for the best scores may represent a difference of 15 raw score points, while 5 percentile points for average scores may represent a difference of only 4 raw score points. This difference in meaning of percentile units arises from the well-known fact that the number of men making extremely high or extremely low scores is much smaller than the number of men who make more moderate scores. Thus, since percentile units are not uniform, and for this reason cannot be averaged, percentile scores have limited value. This problem is discussed further in paragraph 74.

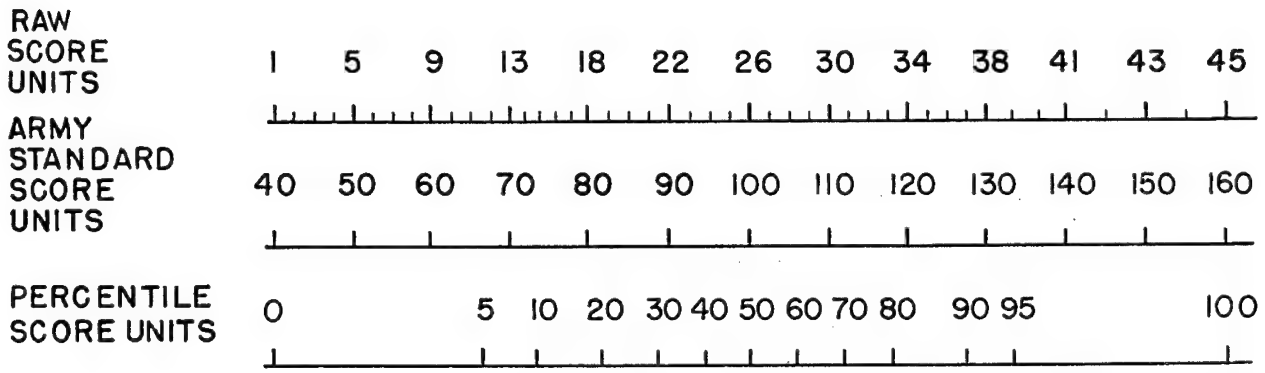


Figure 9. Comparison of raw score units, percentile score units, and Army standard score units for a distribution of test scores with raw score mean of 26 and standard deviation of 8.

Section III. STANDARD SCORES

70. General

a. A more flexible type of standard than any of the types mentioned is the standard score system. The standard score system retains most of the advantages of the percentile score with the added benefit of equal score units. Figure 9 shows how raw score units and percentile units may differ in value, whereas, standard score units are of uniform value throughout the scale. A standard score shows more precisely how much better or poorer than the average score any particular score is.

b. Every classification specialist should be thoroughly trained in the statistical techniques involved in computing standard scores and converting them into Army standard scores. These techniques are described briefly in the following paragraphs of this section. More complete and detailed information can be found in any standard textbook in psychological statistics.

c. Readers of this manual, other than classification specialists, may well omit paragraph 72 which describes the computation of the standard deviation and the standard score.

71. Standard Scores Measure Relative Performance

a. There is more to the standard score system than a comparison of the individual with the average performance. Standard scores also compare people with one another. Some tests are easier than others; some groups taking the test are more nearly of the same level of ability or knowledge than are others. In a group whose members are fairly equal in ability, a high raw score represents more of an achievement than in a group whose members are of a wide range of ability. In the first case, a high score stands out from the rest of the group. In the latter case it does not. The way this works can be seen in the following illustration:

Scores on Test A	*Scores on Test B*
80	80
60	77
58	74
56	68
53	46
49	40
44	30
40	25
8 / 440	8 / 440
55 -Mean Score	55 -Mean Score

On Tests A and B the average or mean scores made were both 55. On Test A, however, the person achieving a score of 80 had 20 score points more than the next highest individual, while on Test B a score of 80 was only 3 points more than the next highest score. On Test B there were a number of scores very close to 80. The man achieving 80 on Test A was outstanding on that test. The man achieving 80 on Test B was only slightly superior to a number of other people who took the test.

b. A standard score system would take into consideration not only the mean performance of the group, but the relative performances of all others taking the test by indicating how far from the mean score each individual score is. How this is accomplished will be illustrated by describing the method of computing standard scores.

72. Computation of Standard Scores

a. There are two elements which must be known before the standard scores can be computed:

(1) *The mean score,* which is merely the mean of all the raw scores of the standard reference population (par. 34).

(2) *The standard deviation,* a measure of the spread of scores or how much members of the tested group vary in ability among themselves. The standard deviation is represented as a certain distance on the scale of measurement above and below the mean score made by the people taking the test. It is computed in the following manner:

(*a*) Subtract the mean raw score from each raw score.

(*b*) Square each of the remainders, or deviations, so obtained.

(*c*) Add all of the squared deviations together.

(*d*) Divide this sum by the total number of scores.

(*e*) Extract the square root of this quotient.

b. When the mean raw score and the standard deviation of the raw scores are known, any individual's raw score on a test can be converted to a standard score by the following method: Subtract the mean raw score from the individual raw score and divide by the standard deviation.

$$\text{Standard score} = \frac{\text{Individual score minus mean score}}{\text{Standard deviation}}$$

For any score on any test, therefore, a standard score can be computed and tables can be drawn up showing the standard score equivalents for all the raw scores. It can be seen that the standard score takes into account the man's deviation from the mean and the relative amount of variation in the group as a whole.

73. The Army Standard Score

There are two undesirable features of standard scores which have led the Army to make a slight modification in their use. For one, the standard score scale is short; a range from —3 to +3 takes in practically all cases. In order to make a fine enough differentiation between men, it is necessary to resort to the inconvenience of decimal scores. Another disadvantage is that all scores below average are negative in sign, another inconvenience. The Army, therefore, multiplies each standard score by 20 in order to get rid of the decimals, and adds 100 to each score to get rid of the negative sign. If a man gets a score which is exactly the same as the mean score on a test, his standard score is equal to 0 (computed as explained in par. 72 above), and his Army standard score is equal to $(0 \times 20) + 100 = 100$. Thus, the Army standard score system has an average of 100 and standard deviation equal to 20 points. The possible scores that can be obtained on an Army standard score scale range from approximately 40 to 160. It should be realized that the selection of 100 to represent

the mean and 20 the standard deviation is purely a matter of convenience. Other numbers could have been used without altering the relationships of the various scores to each other and to the mean.

74. Interpretation of Army Standard Scores

a. Normal Distribution of Human Abilities. In order to interpret Army standard scores, it is necessary to understand a little of the nature of the distribution of human abilities. If the various raw scores made by all persons taking a test are plotted on a graph, the resulting distribution of scores looks something like figure 10.

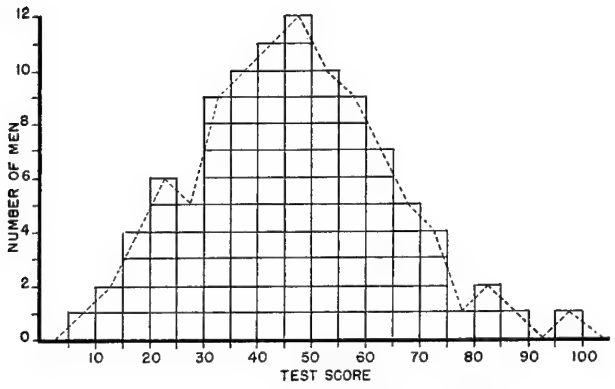


Figure 10. Graph of a distribution of test scores.

Notice that relatively few people receive either the extremely high scores or the extremely low scores, but that most people's scores tend to pile up in the center. The statement of the distribution of scores on a test may be generalized to include most human abilities. That is, a few people possess a great deal of an ability, a few others very little, while most people possess moderate or "average" amounts of an ability. The distribution of most human traits and abilities has a shape very similar to figure 11. This curve is usually referred to as the normal distribution curve. When an Army instrument is standardized, it is usually given to a large group of individuals representing the population with which it will be used. This group is the standard reference population. The distribution of scores for this group will, in most cases, follow closely the normal curve. The normal curve can then be assumed as a basis for setting up the standard score scale and interpreting scores.

b. Usefulness of the Normal Distribution Curve. Because the distribution of most human abilities follows closely the pattern of the normal curve, this curve is very useful for interpreting scores.

(1) The central point on the baseline represents the mean or average score. All other points along the baseline represent scores of different magnitudes from low to high. It can be seen that these scores can also be conceived as differing by various amounts from the mean score. The standard deviation (par. 72), is the yardstick by which the relative values of the scores may be indicated. It is useful to note the number of people who score at levels of 1, 2, and 3 standard deviations above and below the average score. With a knowledge of the average score and the standard deviation of the distribution of scores, interpretations of any individual's performance on a test can be made and individuals can be compared directly.

(2) In addition to comparing people directly with one another, standard deviations give us much the same information that percentile scores yield—that is, they tell us what proportion of the population which took the test fall above or below any one score. When the distribution of scores on a test follows the normal curve, a score of one standard deviation above the mean score indicates that the individual achieving that score has done better than 84% of all those taking the test. Since all Army standard scores have a mean of 100 and a standard deviation of 20, the individual who receives a standard score of 120 is one standard deviation above the mean (100 + 20) and has exceeded the performance of 84% of the people. If his score had been expressed in percentile terms, he would have had a percentile score of 84.

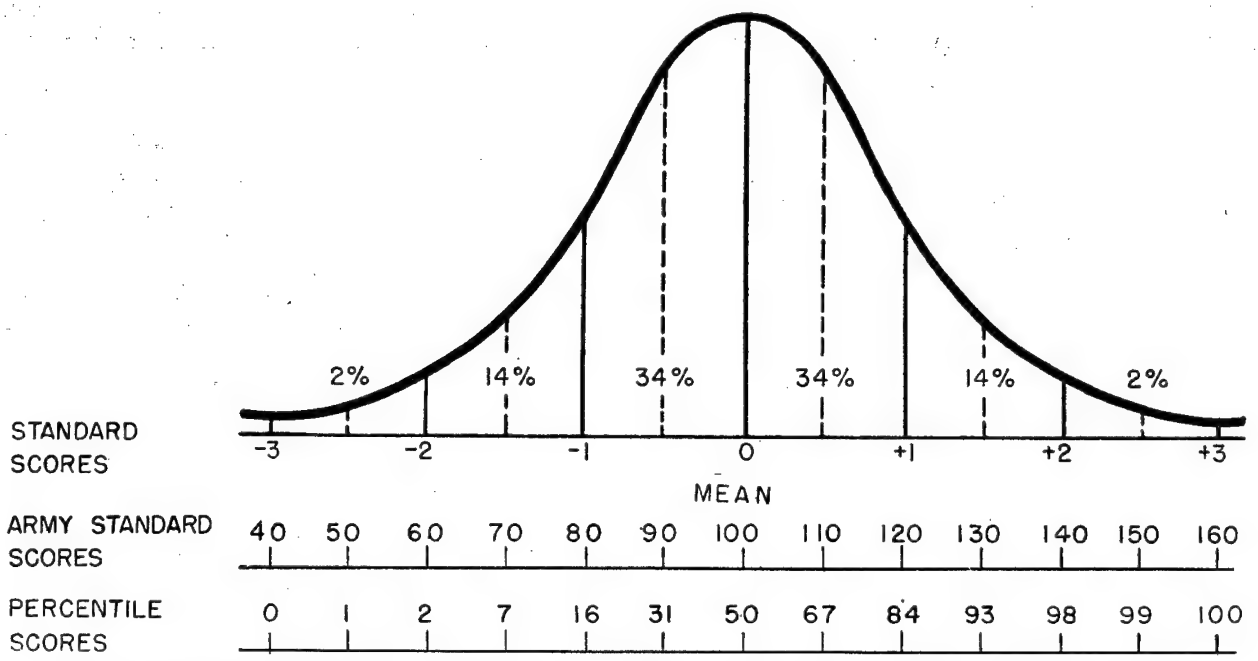


Figure 11. Standard scores, Army standard scores, and percentile scores in relation to the normal curve of distribution.

Each person's score can be compared in this way with those of the rest of the group. Standard scores have essentially the same meaning and the same interpretation from test to test.

(3) Figure 11 shows the relations among standard scores, Army standard scores, and percentile scores and relates them to the curve of normal distribution. It should be pointed out that, once the curve of distribution of scores is known, standard scores and percentile scores may be interpreted in exactly the same way—that is, there is an equivalent percentile score for every standard score. Standard scores, then, yield the same information that percentile scores do, and have the great advantage that since the units are equal, standard scores can be combined or compared with other standard scores. And from the technical point of view, the equal units are of great importance. Without them, statistical computations would be seriously handicapped.

75. Normalized Standard Scores

a. Except for special purposes, a test would not be used if its distribution of scores differed so markedly from the normal distribution as to show, for instance, a sharp peak at one end, or, to take another instance, several widely separated peaks. Abundant experience has shown that in the usual populations, well constructed instruments yield distributions of scores approximating the normal. This experience supports sound theoretical considerations that the traits being measured by the scores are normally distributed in normal populations. There is, then, reason for expecting most test scores to be distributed normally and it is therefore desirable to convert obtained distributions of scores to normal distributions when justifiable. Instances where the distributions obviously should not be normalized are when the number of men taking the test is small or when the men taking the test are highly selected for intelligence, length of military service, grade, etc.

b. Raw scores are normalized by conversion to percentile scores which, in turn, are converted to their equivalent standard scores. Normalized standard scores are thus merely

standard scores based upon a distribution of raw scores that has been made normal. The distribution is normalized only when there is reason to believe that the underlying trait measured is normally distributed.

76. Conversion Tables

The Army provides tables with its tests so that the raw scores can readily be converted into Army standard scores. Each of these tables is based upon the distribution of scores achieved on the test by the standard reference population (par. 34). Thus, any one individual's score is compared with a very large and representative group of Army men, and each individual's standing on the particular trait tested is given in terms of the Army as a whole. All individuals are compared with the same standards. Army standard scores, obtained by use of authorized conversion tables, are therefore much more useful in classification than would be such scores based upon data collected at any single Army installation. Moreover, the use of these conversion tables insures uniformity in translating raw scores into standard scores. The usual table involves two columns of numbers. The first is a list of all possible raw scores; the second contains the standard score corresponding to each raw score. Using the table involves looking up each obtained raw score, then reading off and recording the corresponding standard score.

77. Advantages of Army Standard Scores

Army standard scores are useful for classification and assignment purposes for the following reasons:

a. They state test performance in such a way that small individual differences in ability or achievement are clearly revealed.

b. They tell how a man ranks in comparison with other Army men.

c. They make it possible to compare an individual's expected performance with that of others, and to compare each man's performance on a number of tests.

d. They are mathematically convenient, and therefore make further statistical analysis of data more practicable.

78. Army Standard Scores Are NOT "IQ's"

Army standard scores bear no direct relationship to such concepts as the "IQ" (intelligence quotient), or "MA" (mental age), and Army test results must not be interpreted in terms of these concepts.

Section IV. RELIABILITY

79. Definition

a. An ideal measuring instrument is one which can be depended upon to measure a given object in exactly the same way every time it is used. If a yardstick measured the length of a room as 15¾ feet every time a measurement were made, the yardstick would be a perfectly reliable instrument. But while this ideal is often approached in measurement, it is seldom, if ever, attained. Measurements of a room with a yardstick may vary by several inches if the thumb were used to mark the point where the rule is to be relaid each time. If a cloth tape were used, variation in the measurements might also be found if the tape were stretched more one time than another.

b. With psychological measuring instruments the situation is somewhat aggravated. Tests themselves often fall farther short of perfection than is usual with physical instruments. And, while the dimensions of a room remain quite constant, human beings are changeable and are affected by many conditions which have no effect upon physical objects.

80. Importance of Reliability in Psychological Measurement

a. If the measurements resulting from a test are to be depended upon in making important decisions, they must be sufficiently reliable; there must be evidence that men, repeating the test under the same conditions and without changing significantly, will make nearly the same scores the second or third time as the first. A person's arithmetical ability cannot be represented by a score of 30 in the morning

and 90 in the afternoon, any more than his height can be 5′ 2″ in the morning and 6′ 2″ in the afternoon. People don't grow that fast either mentally or physically.

b. A test's reliability is important also because of its relation to the test's validity. A test can be reliable without being valid—that is, it can be a consistent measure of something other than the particular trait you wish to measure; but a test cannot be valid without being reliable.

81. Conditions Affecting Reliability

There are four main types of conditions which affect the reliability of an instrument—

a. The Instrument. If the instrument is too short, that is, contains too few items, the scores are likely to be greatly affected by chance. If the items are poorly written so that the men have to guess at the meanings, their answers will reflect their guesses.

b. The Men. If the instrument is so long that the men get tired or bored, the responses to the later items will be affected and the test as a whole will not measure consistently whatever it is supposed to measure throughout its length. If some of the men have shown no interest, or if they were tired or in poor health, or if some of them have been specially coached, the consistency with which the instrument measures will be affected.

c. Administration. If the examiner varies his instructions from time to time, if he gives special instructions to some men but not to others, if his instructions are different from those of other examiners, the reliability of the instrument will obviously be affected. If the physical conditions of the room in which the test is being given are poor, the reliability may be reduced.

d. Scoring. Irregularities in scoring affect the reliability. However, in the Army, objective tests are used and the scoring is obviously objective. The evaluation of essay answers, as is well known, is much less consistent than the scoring of objective tests.

82. Estimating the Reliability of a Test

a. General. For two important reasons it is necessary to know just how reliable a test is prior to its use: (1) to determine whether it is suitable for use, and (2) to know just how much unreliability to allow for in interpreting the test results. Before being released, Army tests are subject to rigorous and exhaustive statistical checks to determine the dependability of the measures which may be obtained with them.

b. Method. Reliability may be measured by comparing the scores achieved by men who take the same test two or more times under identical circumstances or who take equivalent forms of the test. If each man gets the same score—or almost the same score—each time, the test must certainly be reliable. But if many individuals get widely different scores each time, the test is unreliable, and no confidence can be placed in any single score. It is almost impossible, however, to give the test more than once under identical circumstances. Besides the obvious effects of familiarity with the test, there are bound to be other changes in the examinees from time to time, and it is often too difficult to construct an exactly equivalent form of the test. In practice, therefore, statistical techniques are used which estimate the result that would be obtained if two equivalent forms of the test had been administered at a single session. The techniques are accurate for most types of tests and have the important advantage of saving both time and effort. The mathematical statement of the result is called the *reliability coefficient.*

83. Uses of the Reliability Coefficient

The reliability coefficient gives important information about the test and the interpretation of test scores.

a. Improving Test Reliability. The first use of the reliability coefficient is to determine whether the test is reliable enough for classification purposes. If the reliability proves to be too low, it can usually be increased by the addition of more items of the same kind as are in the test (par. 22). But since long tests are time-consuming, reliability beyond practical usefulness is not sought. Army tests before being released for use are made sufficiently reliable for use in classification and assignment,

and they will usually remain so provided they are used as directed.

b. Reliance on Test Scores. As pointed out in chapter 1, some calculated risk is involved in personnel actions. Army tests are reliable, but they are not perfectly reliable. A few points difference in scores does not always mean a real difference between the individuals receiving the scores. Yet, when a minimum qualifying score is set, it is recognized and used throughout the Army as the dividing line between those who are acceptable and those who are not acceptable for a certain type of assignment. Admittedly, some of the men accepted would, if retested, fall below the minimum qualifying score and some of those rejected would reach or exceed it. This risk is taken when a decision is made to use a certain point on the measurement scale as a standard for acceptance or rejection. It is a decision essential to an effective classification system. The misclassification of a few men—and their number can be forecast—is weighed against the uncertainty of classification without such a standard. When the minimum qualifying score is disregarded, or when scores are juggled so as to nullify its effect, just that much more of uncertainty—of uncalculated risk—enters into the classification and assignment of men in the Army.

c. Retesting. Because Army tests are reliable it cannot be expected that retesting will, on the whole, result in marked score increases. Familiarity with the test and the test situation may contribute a few points of increase, but experience proves that this increase, on the average, will be small. Moreover, it is important to recognize that there is no virtue in getting high scores for their own sake. The only purpose in using tests at all is to enable predictions of job or training success, and there are seldom grounds for supposing the higher retest score to be a better predictor than the original. On the contrary, there usually is reason to suspect its validity. Therefore, retesting should be practiced sparingly. It is justified only when the record shows a discrepancy which seems to merit investigation. A man's score on a test may be markedly inconsistent with his abilities as inferred from other information (education and occupation, for example). Or a man's records may show obvious discrepancies as a result of errors in recording scores. In such cases, steps should be taken to see that a correct score is obtained and recorded.

Section V. SUMMARY

84. Raw Scores Lack Meaning

a. Army standard scores offer a meaningful way of describing test performance. They compare the performance of the individual with that of other Army men on the particular skill or aptitude which the test measures.

b. The Army provides *conversion tables* which enable field personnel to translate raw scores into standard scores.

c. The *reliability* or consistency of a test as a measuring instrument is carefully checked prior to the test's use. Field personnel can contribute much to maintaining the reliability of a test by strict adherence to regulations governing its use.

CHAPTER 5
THE PRACTICAL VALUE OF SCORES

Section I. THE PRACTICAL SIGNIFICANCE OF SCORES

85. General

As indicated in the preceding chapter, a raw score on any personnel measuring instrument has no meaning by itself. It begins to acquire meaning when it has been converted to a standard score. However, even the standard score does not represent the full significance of a test score. That a soldier has a standard score which puts him in the top 10 percent of all men in some particular respect is factual information. But it will remain an isolated useless fact until its practical significance is determined. In general, personnel measuring instruments are only a means to an end. Their main justification in the Army is the degree to which they can be used to insure that a man assigned to a given job can be expected to perform well in that job.

86. Ambiguity of Instrument Titles

The name of an instrument tells only in a general way what the instrument measures. "Clerical aptitude" tests measure abilities related to proficiency in clerical work and the "automotive information" test gives indication of the probable proficiency of an automobile mechanic. These titles, however, give only rough indication of what the tests are supposed to measure and predict. Full knowledge of how the tests can be used comes only from continued research in which relationship of test scores to performance in a considerable variety of jobs is studied.

87. Factors Affecting the Use of Scores

It is the purpose of this chapter to consider the various factors or concepts which bear upon the most profitable use of test scores. One of these factors, reliability, was discussed in chapter 4. Others are validity, the selection ratio, minimum qualifying scores, and the supply and demand for persons with various skills, knowledges, and aptitudes. General considerations having to do with sampling and the statistical significance of personnel research results will also be discussed.

Section II. VALIDITY AND SCORES

88. The Correlation Coefficient

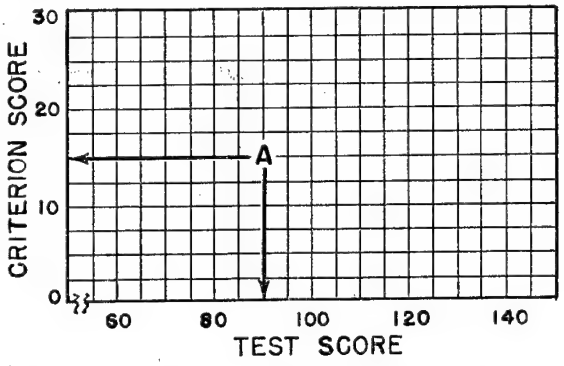
a. General. The correlation coefficient is a statistical measure which, because of it usefulness, is encountered frequently in personnel research. It may be defined as a number, between the limiting values of +1.00 and —1.00, which expresses the degree of relationship between two variables. These variables may be two sets of test scores on a group of men, test scores and criterion measures, scores on the same instrument from two different periods of time, or measures on any other traits, qualities, or characteristics which exhibit differences in magnitude.

b. Graphic Illustration of Correlation. A correlation coefficient can be illustrated simply by means of a graph. In the example chosen (fig. 12), the two variables compared are a test and a criterion, for each of which scores are available for 5 men. Scores of the 5 men are given in the first box of the figure. In the second box the manner of plotting the point representing the paired values for the first man is shown. The horizontal axis is chosen here

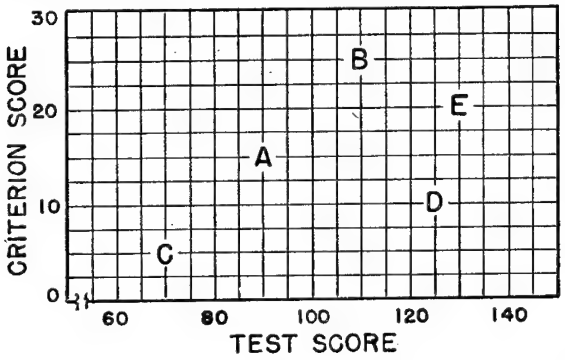
1. TEST AND CRITERION SCORES

MAN	TEST SCORE	CRITERION SCORE
A	90	15
B	110	25
C	70	5
D	125	10
E	130	20

2. PLOTTING THE PAIR OF SCORES FOR MAN "A" ON THE CORRELATION SCATTER DIAGRAM



3. COMPLETE CORRELATION SCATTER DIAGRAM FOR THE EXAMPLE



4.

CORRELATION COEFFICIENT*
(ALSO CALLED VALIDITY COEFFICIENT)
FOR THIS EXAMPLE = .54

*COMPUTING STEPS NOT SHOWN

Figure 12. Illustration of the relationship between test and criterion scores for 5 men.

to represent test scores and the vertical axis the criterion. In the third box, all 5 points have been plotted for the example, and in the fourth box the result of computation (not shown) is given—the correlation coefficient of +.54 for these values. Plotting of such a graph is similar to plotting a distribution of scores, but since there are two scores for each person, the distribution may be termed a double distribution (also called "scatter diagram" or "scatterplot"). As can be seen, it closely resembles the plotting of points on a map through the use of coordinates.

c. Interpretation. If there is no relationship between the two variables, the correlation coefficient is zero. If exact or perfect relationship between the variables exists, a coefficient of either +1.00 or —1.00 is obtained; if high scores on one variable are associated with high scores on the other while low scores on one are associated with low scores on the other, the coefficient is positive; if high scores on one variable are associated with low on the other while low scores on one are associated with high on the other, a negative coefficient is obtained. Figure 13 shows the kinds of graphs associated with correlation coefficients of .00, +1.00, —1.00, and +.55.

(1) The correlation coefficient should be interpreted in terms of a "line of relationship." Such a line has been drawn in the first and fourth graphs of figure 13. The closer the points to this line, the higher the correlation between the two variables. In the second and third graphs of figure 13, where correlation is perfect, all points

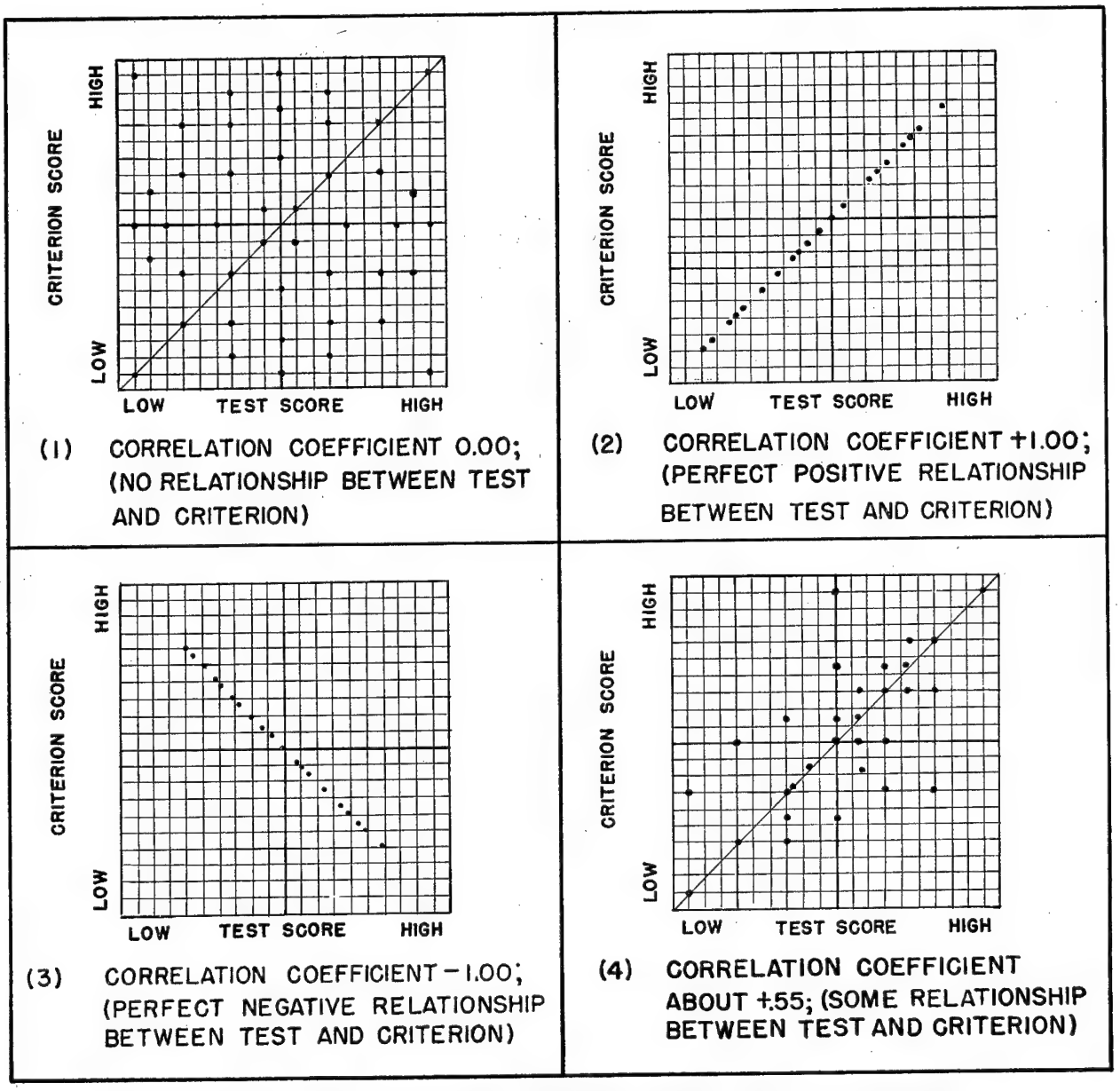


Figure 13. Correlation scatter diagrams illustrating different degrees of relationship.

would fall exactly on this line, and it has, therefore, not been drawn in.

(2) Accuracy of prediction of one variable from another is a function of the degree of relationship between them. In the case of a correlation coefficient of 1.00, prediction is perfect; knowing a score on one variable, one can predict the score of the same individual on the other variable without error. When the correlation is zero, it is impossible to predict one score from knowledge of another.

(3) Any thorough-going discussion of the correlation coefficient should introduce the concept of regression, which may be defined as the tendency of a predicted value to be nearer the average than is the value from which the prediction is made. The lower the correlation, the greater is this tendency. When there is no correlation between two variables, the best estimate that can be made of the value of any variable is the average score on that variable. For fuller discussion of these

ideas, reference should be made to a statistics textbook.

(4) Correlation coefficients are sometimes given special names according to the purpose for which they are used. A correlation coefficient used to express validity is called a validity coefficient; used to express reliability it is called a reliability coefficient.

89. The Validity Coefficient

a. Meaning. The validity coefficient for a test is the coefficient of correlation between the scores on the test and the scores on a criterion representing what the test is supposed to measure. It indicates the extent to which the individuals who score high on the test also tend to score high on the criterion, and the extent to which individuals who score low on the test also tend to score low on the criterion. For example, a very high validity coefficient, say .90, would mean a high degree of relationship between the test and performance on the criterion, such that a man with a high test score would be extremely likely to exhibit excellent performance on the criterion; there would be some chance of error in estimating his criterion performance from his test score, but it would be slight. Similarly, men with test scores at other points on the scale would be very likely to perform on the criterion measure at the corresponding levels. On the other hand, if the correlation between test and criterion is low or close to zero, a man with a given test score might achieve criterion performance anywhere along the scale—high, low, or average.

b. Use With Army Tests. When scores on Army tests are correlated with later performance, validity coefficients of .30, .40, and .50 are often obtained. Considerable assistance in prediction can be obtained from use of such tests, although it is obvious that the higher the validity coefficient, the greater this assistance would be. Much of personnel research is directed toward exactly this problem—obtaining tests which will predict job or training performance with the highest possible accuracy. Even in cases where research has as yet been able to produce predictor instruments of only low validity, their contribution to selection and classification may be a sufficient improvement over purely random placement of men to make their use worthwhile. This consideration leads to an important point in the interpretation of validity coefficients—the size of coefficient which is deemed useful is relative to the type of instrument and the situation in which it is to be applied. In some situations, a predictor with very low validity may be far better than none at all; in other situations the validity must be much higher before the instrument is worth introducing.

c. Specificity of Validity. From the foregoing it can be seen that the validity coefficient shows the degree of relationship between scores on a test and performance in some particular job. Validity, then, has meaning only in relation to a particular test and a particular job. A test may have high validity in predicting performance on some jobs, moderate validity for certain other jobs, and no validity at all for certain additional jobs. It may have validity in predicting success in training but little or no validity in predicting success on the job for which the training is conducted. It is meaningless to speak of *the* validity of a test; it is necessary to specify what it is valid for.

d. Use With Test Batteries. So far, the discussion of validity has been limited to the predictive value of scores from a single test. Actually, as the discussion in chapter 2 indicated, several test scores are often weighted and combined to give a single composite score for the battery. Performance on a job can often be much better predicted if measures of a variety of aptitudes and traits are used in estimating later job performance. The discussion of the validity of a test score applies equally well to the validity of a composite score resulting from a combination of several test scores.

e. Expectancy Charts. In the use of a test for selecting men who will perform successfully on a given job, the meaning of the validity coefficient is sometimes expressed in the form of an expectancy chart. Such a chart would show what the probabilities are that men making various test scores will achieve better than average success on the particular job. The validity coefficient can thus be translated into terms easy to apply in a practical situation.

90. Face Validity

a. The foregoing discussion of validity is based on the empirical determination of degree of relationship between a set of test scores and a set of criterion scores. Empirical validity is expressed as a validity coefficient computed from the two sets of scores. Face validity, on the other hand, is not a computed validity. The term "face validity" is actually used in several different senses and it is important to keep the several meanings clearly in mind. However, they all have one characteristic in common—they all involve a guessed or inferred validity rather than an empirical or computed validity.

b. First, face validity may refer to the overall appearance of the content of a test—whether or not the items are obviously pertinent to the subject of the test. Face validity in achievement test questions might mean that the questions in an automotive test are only those that bear directly on facts and skills in the automotive area. Face validity in rating scales refers to use of items and rating categories which relate directly to the subject of the rating. Face validity in these examples is directed at obtaining general acceptability of the instrument among those who will have contact with it; the user of the test, not having access to empirical validity data, may react unfavorably to an item which does not seem "on the face of it" to be relevant to the subject.

c. A second use of the term "face validity" occurs where a new test is constructed on the basis of prior knowledge; items are selected from other tests on the basis of statistics computed in previous investigations. In the absence of contrary information the validity of the items in the new test may be inferred.

d. A third use of the term "face validity" occurs when subject matter experts determine the items to be included in a test in terms of requirements of the job for which the test is being constructed. Such instances arise in the construction of achievement tests where questions are written to cover essential aspects of each job. The subject matter expert provides items which in his judgment are likely to be answered correctly by men who know the subject and incorrectly by those who don't know the subject. This type of face validity is also referred to in chapter 7 as "built-in" validity.

e. A final point—face validity and empirical validity are both to be distinguished from internal consistency, which refers to the correlation between each item and the other items in the test (ch. 2). Internal consistency may be determined either by subjective estimate or by statistical computation from empirical data, but it is usually not thought of in terms of test validity

f. It should be clear that face validity is no general substitute for empirical validity. In the case of measurement instruments which involve right and wrong answers (achievement tests, for example), the "built-in" face validity may be defensible. However, for items in instruments which do not involve right and wrong answers (for example, self-description forms) empirical validity must be determined. As a general rule, attempts are made to determine empirical validity for all types of instruments. Sometimes face valid items are found to have little or no empirical validity; and conversely, items with little or no face validity are found to have empirical validity.

Section III. SELECTION RATIO

91. Definition of Selection Ratio

While the validity coefficient is most important in deciding whether a test score will be useful in selecting men for some job or training course, other factors are also important. The selection ratio is one such additional factor. Specifically, the selection ratio is the number of men needed (demand) divided by the total number of men available (supply).

92. Significance of the Ratio

In general, if the selection ratio is small, that is, if there is need to select only a few from a large number of available men, test scores are highly useful. If this is the case, the cream of the crop can be selected. If, on the other hand, the need exceeds the number of men available, tests will be of no value whatever since all available men will have to be used. It will be

seen below (par. 94) that the selection ratio has an important bearing on the establishment of minimum qualifying scores.

93. Application in the Army

a. The selection ratio is a highly important factor in determining the quality of men finally selected for some given job. A measuring instrument does not have a high or low validity of and by itself. A low numerical value of the validity coefficient may actually indicate very useful validity if the selection ratio is small. To take an extreme case, an instrument with a validity coefficient of .15 may be very effective in selecting men who will be successful on the job if only 10 men out of 1,000 available are to be selected.

b. The Army uses the basic ideas underlying the selection ratio to good advantage in selecting men for important assignments. A small selection ratio is used in selecting officer candidates. Selection of candidates for the United States Military Academy is a second instance in which the factor plays an important role. In jobs less important and involving less responsibility, lower standards and larger selection ratios are employed. Of course, it is not always possible to control the selection ratio and keep it stable since the manpower picture changes from time to time. This means that careful staff planning is essential to obtain reasonably accurate estimates of the demand for manpower and the supply—an important illustration of the close relation that exists between Army personnel policy and Army personnel research.

Section IV. MINIMUM QUALIFYING SCORES

94. Factors in Determining Minimum Qualifying Scores

In order that assignments will be made effectively on a uniform basis throughout the Army, minimum qualifying scores are set for almost all possible assignments. It can be seen from the discussion of the selection ratio that the particular standard score designated as the minimum acceptable score is to a considerable degree determined by the importance of the job, the number of men who are available, and the number who are needed. Generally, then, a minimum qualifying score for a job is set by considering the distribution of scores for the test or battery found to be most valid for that job and deciding on the portion of those available needed to fill the vacancies that exist.

95. Conflicting Demands for Men

The minimum qualifying score cannot always be set for one Army job without taking into account the needs for other possible jobs. When feasible, the more important assignments should be filled first. Thus, while high minimum qualifying scores will usually be used when the job is important, less important jobs must be filled from those remaining. Often, fortunately, the type of man needed for one job is not the same as that needed for another. In selecting clerks, for example, most of the men selected would not be expected to do well in various jobs calling for mechanical skill and ingenuity.

96. Minimum Standards of Performance

While minimum qualifying scores are usually set as a function of the expected supply in relation to the needs for a given job, an additional factor is sometimes important. Some men are so poor for some jobs that they would contribute nothing or make errors costly enough to over-balance what they did accomplish. In other words, minimum standards must be used to avoid placing men in jobs where their probable performance is such that they will be a detriment rather than an asset. Generally, minimum qualifying scores should be set as high as the supply-demand situation allows. They should also be sufficiently high to insure that minimum standards of performance are met. Caution must be exercised that the minimum standards are not set arbitrarily high—they should be based on careful study and not on mere desire to select nothing but the best.

97. Need for Uniform Standards

If a minimum qualifying score set on an Army-wide basis is not adhered to, confusion will result. If one installation uses a minimum

qualifying score of 130 for Officer Candidate School while another uses a score of 110, men scoring between 110 and 120 will be admitted to Officer Candidate School from the second installation, while men with scores of 128 are being rejected in the first installation. The quality of the men entering the school will depend on the installation they came from, rather than on the instruments. Not only would the quality be uneven, but it would make it extremely difficult to maintain an effective selection system.

Section V. SOME GENERAL CONSIDERATIONS INVOLVING STATISTICAL SIGNIFICANCE

98. Necessity for Sampling

a. Data collected for study in personnel research—whether it be test scores, course grades, ratings, or personal history information—usually represent only a portion of all of the data available. In most cases, it would be expensive, difficult, and unnecessary to utilize a total population when a sample adequately representative of the total could be drawn and used. For example, if the average score on a new test for the Army population is to be determined, it *might* be possible to administer it to every soldier and to compute the mean of this tremendously large number of scores. However, such a procedure would obviously be time-consuming and costly; practical considerations would make it impossible as well. It usually suffices to obtain a sample of men from the total population and to use the mean of their scores as the mean that would be obtained in the total population.

b. Reasons other than time and expense sometimes make sampling necessary. For example, if an arsenal manufacturing hand grenades wants to be sure that they will explode properly, an obvious test would be to explode them. Equally obvious is the fact that after this test there would be no grenades left! In this type of testing situation, it is necessary to select a sample and to apply the test to the sample only.

99. Errors Due to Sampling

Whenever a sample is drawn from a total population, it is to be expected that any statistic, such as a mean or a percentage, that is based on that sample will vary from the "true" value of that statistic. A "true" value here is used to mean the value that would be obtained from results based on the entire population. Consider again the example of computing a statistic such as the mean test score for the Army population. A sample of men might be drawn and the mean of their test scores computed. Then another sample of men might be drawn, independently of the first selection, and the mean of their test scores determined. If this process were to be repeated many times, the mean from the successive samples would be observed to fluctuate. Most of them would fall at, or very close to, the true mean, but a few would, by chance, deviate more. The likelihood of obtaining a mean very much different from the true mean would be small, but it might occur in one of a large number of samples. If a large number of these means were obtained, they would be found to be distributed around the true mean according to the normal probability law. These successive means (obtained by sampling) could themselves be treated like scores, and the average and standard deviation of the means computed; about 68 percent of them would be found to lie within one standard deviation on either side of their average mean. This standard deviation of sample means is not the same as the standard deviation of the original test scores; it is referred to as the standard error of the mean σ_m. Estimates of the standard error of a statistic usually can be made from statistics based on one sample. The standard error of the mean, thus, can be shown to be equivalent to the standard deviation of the individual scores from one sample divided by the square root of the number of scores. Discussion of standard errors of other statistics can be found in any statistics textbook.

100. Factors Affecting the Adequacy of a Sample

a. Manner of Selecting a Sample. First and foremost, statistical results based upon a sample will be affected by the manner in which the sample is selected. Bias entering into selection of a sample will be reflected in results based on that sample; the extent to which statistics derived from a biased sample can be generalized to the population is usually indeterminate. Discussion here refers to the situation where samples were randomly drawn from the total population. The conditions for simple random sampling are that each individual in the sample has an equal chance for being drawn and that the selection of one individual shall in no way affect the selection of another.

b. Size of Sample. Statistics based on large samples are more likely to represent population values than those based on small samples. For example, if only two men were selected from the total Army population, the chance of a large difference between the average of their scores on a test and the total Army average would be great; however, if 2,000 men were randomly selected, the chance of an equally large difference would be very, very small.

101. Statistical Results as Estimates

Whenever any statistic, such as a mean, is computed on the basis of a sample, the question arises—how good an estimate is this statistic of the value for the entire population? Answers to this question can be made in terms of the "standard error," described in paragraph 99 above. If the mean of a sample of 100 test scores is found to be 50, and the standard error of the mean is 1.5, these data may be interpreted as follows: If a number of random samples of 100 cases each were to be drawn from the original population, and the mean test score computed for each sample, 68 percent of these means might be expected to lie within one standard error of the true or population mean, 98 percent within two standard errors. This may be stated differently: There would be a 68 percent chance that the interval between 48.5 and 51.5 (50 ± 1.5) would contain the population mean, a 98 percent chance that the interval between 47 and 53 ($50 \pm 2 \times 1.5$) would contain this value. Thus, "confidence limits" can be set on the variation that can be expected in results based on a sample randomly selected from a given population. From this information, inferences can be made, with stated degrees of confidence, as to the true values of various computed measures, or the values that would obtain if the entire population were used. It should be noted that the type of error described here is a chance error, a fluctuation due to sampling. It cannot be eliminated; it is to be expected whenever a sample is used in place of an entire population. It is important that errors in a sample be restricted to such chance errors, and that they not be due to definite biases which cause the sample to be unrepresentative of the total population.

102. Test Scores as Estimates

In basing decisions on a man's test score, a practical question arises—how much meaning can be attached to a given score? If the man were to be retested a number of times, what fluctuations might occur in his score? How likely is a given test score of 105 to become 106 on retesting, or 103, or even 120? For what reasons might these variations occur? Fluctuations of this type, although they are called "errors," are somewhat different from the fluctuations or errors discussed in paragraph 101. Rather than being due to the effects of simple random sampling as in the previous examples, variations in individual test scores are due to lack of refinement in the measuring instrument. The error of measurement here is of the same type that exists in physical measurements—length can be measured by a yardstick, by a foot-ruler, or by a micrometer, but with each type of measuring stick, successive readings will vary, the degree of variation depending on the refinement of the measurement made. In the personnel testing situation, another way of looking at these errors of measurement is that they are errors of a "mistake" type, where the examiner did not happen to time the test precisely, where unusual distractions might be present, where the examinee's motivation was exceptionally poor, or various similar circumstances existed. The "standard error of measurement" is used to represent the variation in measurements ob-

tained with a particular test. The extent of this error is a function of the reliability of the test as a whole and also of the total expected variation in the test. From this discussion of types of errors of measurement, it should be clear that a man who obtains an Army standard score of 90 on a test might well on another occasion obtain 93, or 89, or some other value, probably within the range of 85 to 95.

103. Significance of Differences

Decision is often necessary in personnel research studies as to whether differences obtained are true differences or could be due to chance errors in the values being compared. Necessity for evaluations of differences may arise in situations such as the following: comparing average grades of successive classes in schools, comparing validity coefficients of different tests in order to select the best predictor, comparing reliabilities of tests, selecting the best method of training for a job by comparing average performance ratings of groups taking the training. For example, consider two groups, A and B, with mean scores of 105 and 115, respectively, on a test. If the standard error of each mean turns out to be 1.5, there is a 98 percent chance that the true mean of group A lies between 102 and 108 ($105 \pm 2\sigma_m$), and that the true mean of group B is somewhere between 112 and 118 ($115 \pm 2\sigma_m$). Since the upper limit of group A and the lower limit of B do not overlap, there is considerable chance that there is a true difference between A and B on this test. On the other hand, if group C obtains a mean score of 107, the same standard error and the same confidence limits would place the true mean of this group somewhere between 104 and 110. The overlap of these limits with those for group A makes clear that there cannot be the same assurance of a difference between A and C that there was with regard to A and B; there is a fair chance that their true means might be identical, although their obtained means were different. Methods have been derived for establishing specific confidence limits with regard to the significance of differences in statistics computed from samples, with results stated in terms of probabilities. Elementary statistics textbooks in the field of psychology should be consulted for further development of these ideas and for explicit formulas. It should be kept in mind that if a particular statistical comparison indicates no significant difference, it does not follow that the true measures are identical. It may well be that there is a true difference which the samples drawn failed to reflect on this occasion.

104. Practical Significance as Compared with Statistical Significance

Test scores, as well as research results, must be interpreted not only in terms of their significance as computed by the best statistical tools available, but also in terms of elements in the real situation in which they are being used. Thus, sometimes a difference between two values which has been shown to be "statistically significant" may have to be ignored on practical grounds. For example, average performance rating of men who have undergone training method A may be superior to that of men under method B; the probability may be almost zero that the difference could have arisen by chance; yet, factors such as cost, time, convenience, or availability of equipment may be such as to more than offset the gain in performance. A test may be built to yield extremely refined differentiations in possible scores, with a very small standard error of measurement; yet this refinement of measurement will be useless if all that is needed in a given situation is to lump examinees into two general categories, such as those acceptable for training and those not acceptable. Further, it is sometimes true that the real situation does not permit obtaining data on a sample of men drawn from the population according to desired sampling plans; then allowance must be made for this in the interpretation of research results based on the data. Statistical results must always be applied in terms of the real situation. Statistical techniques are tools to aid in dealing with practical problems and can be used only as such. Their usefulness must be related to the practical aspects of personnel problems.

Section VI. SUMMARY

105. The Interpretation of Test Scores

a. The validity of an instrument must be known before the scores can be properly interpreted. Validity is, however, specific to a given job or assignment; a test may have different validities for different purposes.

b. The selection ratio—the number of applicants needed divided by the number available—is an important consideration. When the selection ratio is small, men with high scores can be selected and a low failure rate on the job is to be expected. When this ratio becomes large, selection is less efficient and the on-the-job performance of those selected approaches what would be obtained if men were selected at random.

c. A minimum qualifying score is used to set Army-wide standards for particular assignments. The number needed by the Army for that assignment and the number available were shown to be the two most important considerations in deciding upon a critical score. The need to avoid selection of men whose performance will probably be completely inadequate is an additional factor in setting minimum qualifying scores. Great care must be exercised in determining what completely inadequate performance is.

CHAPTER 6

USE OF APTITUDE MEASURES IN INITIAL CLASSIFICATION

Section I. INITIAL CLASSIFICATION

106. Purpose

In the course of his first days in the Army, an enlisted man goes through the process of initial classification. At its close, a recommendation has been made as to the military occupation in which he will be assigned. If the task of classifying him has been well done, this will be the job in which, in view of immediate needs, he fits best into the Army's complex manpower structure. This means that it is a job he can do well or learn to do well, that his abilities and skills are not thrown away on the job, that it is a job in which the Army very much needs men like him at the time. How can this decision, a decision of continuing importance to the Army and to the man throughout his Army career, be made intelligently on the strength of the brief and formal routine that is initial classification?

107. Basis for Initial Classification

a. Before the choice of assignment for an enlisted man is made, two kinds of inventories are necessary. First, there must be an array of facts about jobs in the Army—What jobs have been set up as Army jobs? How many jobs of each type are to be filled at this time? Which jobs must be filled immediately? How many men have to be trained for the various jobs within a certain time? In regard to each job, such questions as these must be answered—What kind of physical effort must a man be capable of to do it? What skills must he have or will he have to acquire? How much will he have to learn? How difficult are the problems he will have to solve?

b. The man, too, must be inventoried. This inventory has one main objective—to obtain advance estimates of his probable success in different kinds of Army jobs. The classification officer who finally makes a recommendation for the man's assignment will have at hand all the information that has been collected about the man in his few days of processing. The classification officer knows what the man's physical status is. He knows how much formal education the man has completed, what games or amusements he likes, what jobs he has held, how long he worked in each one, as well as what kind of work the man thinks he would like to do. And, most important, the classification officer has at hand estimates of how well the man can be expected to do in a number of occupational areas, these in the form of scores on the battery of aptitude tests the man has just taken.

c. The process by which all of this information is obtained from the enlisted man and applied in initial as well as in later phases of classification is described in SR 615–25–25. It is the objective of this chapter to explain how personnel measurement techniques are employed to increase the effectiveness of this classification.

Section II. APTITUDE AREAS

108. What Differential Classification Is

a. Army jobs require varying patterns of skills and abilities. Each man is better at some things than at others. When the skills required in an occupation are among the things a man does best, or can learn to do most easily, the job is likely to be well done. In addition, other

jobs which are of the same level but which require different abilities can be done better by other men whose skills and abilities, attained or potential, fit those jobs. The principles of differential classification are practiced by the Army as the means by which a man's strong and weak points are considered in recommending him for assignment to a certain job or training for that job.

b. Differential classification usually employs results on a battery of tests which provide estimates of a man's probable success in a number of different occupational areas. This makes it possible to see in which fields he is likely to do best and in which fields his prospects of success are relatively poor.

109. Definition of Aptitude Areas

a. The Army Classification Battery. The set of tests used to evaluate the capacities of enlisted men during initial classification is the Army Classification Battery. Each of the tests in the battery is intended to measure a different aptitude or skill important in one or more types of Army jobs. The battery is now made up of a number of comparatively short tests, including tests of reading and vocabulary, arithmetic reasoning, pattern analysis, radio code aptitude, mechanical aptitude, clerical aptitude, shop mechanics, electrical information, radio information, and automotive information.

b. What an Aptitude Area Is. In use, the Army Classification Battery follows a somewhat different pattern from that of the typical selection battery developed and validated as explained in chapter 2. The tests in the Army Classification Battery are not employed as successive hurdles. Nor are they used as a single test composite to provide an over-all measure of the individual. Instead, they are used in a number of different combinations, each made up of two or more tests. The score on each group of tests represents a combination of abilities which is considered important for satisfactory performance in a number of Army jobs. The combinations of tests, together with the group of jobs for which each combination predicts success, are the Aptitude Areas. There are at present ten such combinations of test scores, each associated with varying numbers of Army jobs. For example, an estimate of probable success as a clerk is based on a composite of scores on the reading and vocabulary, arithmetic reasoning, and clerical speed tests. Associated with this same combination of tests are other jobs such as toolroom keeper, tabulating machine operator, and classification specialist, which at first consideration would seem to have little in common. The method of combining tests to predict performance in a job area is described in paragraph 110*c*.

c. Changing Character of the System. In thus defining Aptitude Areas, two points should be emphasized. Both have to do with the developing nature of the system. In the first place, it is recognized that, as research on the tests goes on and as Army jobs change, new tests may profitably be added to the battery or substituted for some tests now in the battery. Secondly, there may be changes in the composition of the Aptitude Areas or shifts in the jobs associated with the various areas. Such changes would follow upon a continuing analysis of Aptitude Areas in operational use. Validation studies are undertaken to check on how well the Aptitude Area scores indicate the relative success in performance or training in the case of specific jobs. At the same time, it is possible to try out as predictors for those jobs combinations of tests other than the one with which a job is now associated. Research is continuing toward the development of tests which are more independent of each other and provide measures of more clearly defined traits. This might mean eventually a battery composed of a greater number of short tests, each of a highly specific ability. The Aptitude Area scores derived from such a set of tests could be expected to overlap each other less than do the present Aptitude Areas. In the present battery there are certain tests, such as the Arithmetic Reasoning Test, which are valid for almost all jobs in the Army. These have been included in a number of the Aptitude Area combinations and account for much of the interrelation between the area scores, which is considered more than is usually desirable with measures for this purpose. Another limitation of the present set of Aptitude Area scores is that there is no representation of personality characteristics, such as interests, ability to work in a team, courage under fire, leadership. Much of what-

ever conclusions are drawn in regard to these personal characteristics depends on the personal judgment of the classification and assignment officer.

110. Development of Aptitude Areas

a. Limitations of a Single Over-All Score in Classification. In the early days of group testing on a wide scale in the Army, main reliance was placed on a general measure of ability, typified by the early forms of the Army General Classification Test (AGCT). Scores on this test were widely used, not only in allocation of men to fill job quotas, but in selection for specialist school and officer candidate training. The test sampled several areas of ability, but only a single over-all estimate of the enlisted man's trainability was obtained. Use of a single measure in classification was adopted in response to the pressing need to fill jobs with men who could do them. The simplest way of getting a job done satisfactorily, it was felt, was to put a man in that job who had at least enough general ability to learn a job of that level. It was understood, however, that in a way this was a wasteful solution. The Army General Classification Test score provided what was, in effect, an average of a man's estimated skills and abilities. Obviously, he must possess some skills at greater than his average level, and others at less. To classify him at his average level was to fail to make use of his better skills. On the other hand, he might almost as readily be assigned to a job which made demands on his poorer skills. In that case, there was a definite probability that the job would not be adequately done.

b. Steps Toward Differential Classification. In a later modification of the Army General Classification Test, an attempt was made to provide a means of considering these variations in level of an individual's capacity in several areas. The new form of the test yielded four part scores—Reading and Vocabulary, Arithmetic Computation, Arithmetic Reasoning, and Pattern Analysis (a measure of ability in perceiving spatial relations). Each of these part scores could be taken as indicative of aptitude for certain types of work. They were also summed to provide a total or over-all score. These measures were supplemented by such tests as the Mechanical Aptitude Test, the Clerical Aptitude Test, and the Army Radio Code Aptitude Test, which had relevance for certain groups of occupational specialties. The groundwork thus was laid for application of the principles of differential classification. The intent was that the classification officer, in selecting a man for assignment, would take into account his varying capacities in the four areas and supplement these measures by the special tests.

(1) In practice, however, he continued to give priority to the general over-all measure. The reason for this lay not so much in the persistence of the habit of using an over-all measure—though this was undoubtedly a factor—as in the cumbersome process that had to be gone through in giving consideration to the four part scores and the total score as well. The classification officer had, in theory at least, to consider each score in relation to all the occupations open at the time. He had to decide how important each score was to the job, what combination of tests could best be taken to indicate prospects of success in that job. If pursued to the logical conclusion, the process was endless. In practice, no classification officer went through all these steps. He took short-cuts—some legitimate and some not. For example, he would then—as he does now—automatically leave out of consideration jobs inconsistent with the enlisted man's educational background. Then, as now, when a man was found to have done a certain type of civilian work satisfactorily, the classification officer would, if possible, assign him to a similar job in the Army. These are examples of good classification practices. A less advantageous practice, and one which worked to nullify the benefits of classification on the basis of the varying capacities within the individual, was to consider only one of the part scores in making an assignment, or to consider only the total score.

(2) Where the system fell short of its purpose was in not providing a means of using, in an organized fashion, the information provided by the several tests scores. A need was also felt, in applying the scores to classification problems, for a finer breakdown of aptitude measures than the four part scores the AGCT provided.

c. Formation of Aptitude Areas. The four parts of the AGCT and the other tests given regularly in the course of reception processing had all been tried out with many different jobs to find whether they provided useful estimates of later performance. As a result of these validation studies, it was discovered that the same test or group of tests turned out to predict likelihood of success for a number of different jobs. The jobs, different one from another as they might be, were organized into occupational areas. This is the basis on which Aptitude Areas were formed. The tests which work together best to select men who can be expected to do well in those jobs make up the test combination which yields the Aptitude Area score. The data employed in defining the job areas and the combination of tests associated with each are the same data previously used in setting up expectancy charts to show what chance a man making a particular score on a given test had of making an average-or-better grade in a particular training course. These expectancy charts appeared in the previous edition of this manual and are not repeated here.

111. How Aptitude Areas Are Used in Classification

a. When the classification officer approaches the problem of evaluating the man and deciding upon his initial assignment in the Army, he has at hand a "profile" showing the man's Aptitude Area scores. This shows at a glance the area or areas in which the man has the best prospects of success. It can be related directly to the types of job he is likely to do best. By considering this information against the jobs he has to fill, the classification officer can usually narrow the choice of assignment down to a reasonable number of alternatives.

b. It should be emphasized here that this selection of initial assignment comes from a synthesis of all the information collected about a man. With different classification problems, different aspects of the body of information come to the fore. In some cases, a man's education or lack of it will determine whether he is assigned for certain types of training. As mentioned earlier (par. 110*b*), in every case where an enlisted man has gained competence in some civilian occupation, this experience outweighs other indications of his probable success. Even though his score in that Aptitude Area indicates that he would require much longer than the average training period to learn a job, if such training has already been effectively accomplished while he was a civilian, the Army can take advantage of it.

c. It will be seen from this discussion that even when Army needs do not interfere, assignment is not always made in the area in which a man makes his highest score. The practice followed, where possible, is to make the assignment in an area in which he made one of his higher scores. The classification officer often finds it necessary to decide between several Aptitude Areas in which the man has made almost the same score. Such factors as physical profile, school quotas, and work experience then assume relatively greater weight in determining recommendation for assignment. Still another factor must be considered, namely, the level of ability required. A man's best score may be in Aptitude Area requiring a high level of ability, and his second best in an Area requiring a lower level of ability. It is possible that he will perform better in the Area requiring a lower level of ability.

112. Gain Resulting From Classification by Aptitude Areas

a. General. Initial classification based on Aptitude Areas provides a systematic approach to estimating likelihood of success in various job areas and makes it less likely that a particular strong point in an individual will be overlooked.

b. Advantages to the Individual. The effect of classification by Aptitude Areas on the individual placement is first considered. Advan-

tages are indicated in figure 14 which shows the Aptitude Area scores of two enlisted men. Both A and B score about 100 on AGCT.

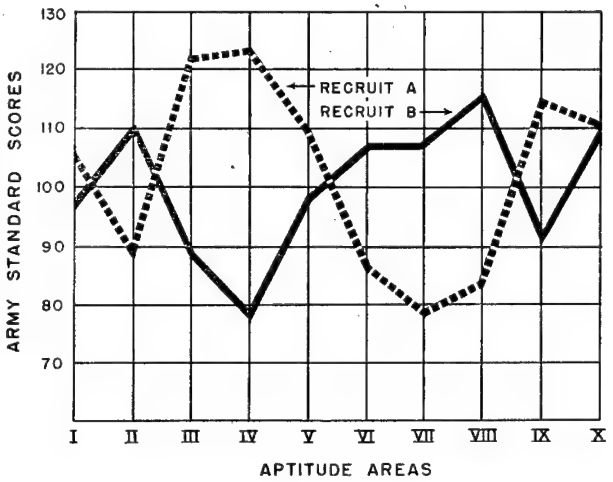


Figure 14. Aptitude area profiles for two recruits.

However, it is clear from the differences in the two profiles shown in the figure that they have very different patterns of skills and aptitudes. Obviously, if classification were based on only one score, such as the AGCT, these two men could be used interchangeably. On such basis, if A were classified for a job predicted by Aptitude Area II or VI or VII or VIII, or B were classified for a job predicted by Aptitude Area III or IV, or IX, they would both be placed in jobs in which their likelihood of success is poor. By the application of Aptitude Areas, both A and B can be placed in one of their best areas on the basis of their measured strengths and weaknesses, rather than having such placement left to chance. In their best areas, A and B can be classified for jobs in which their likelihood of success is relatively higher.

c. Advantages to the Army.

(1) With Aptitude Area classification, the Army can make better use of its men. Emphasis is on putting the man in the Army job he can fill best rather than in placing him in a job he can fill acceptably. Even when requirements of the service prevent a man's assignment in his best Aptitude Area, he can usually be placed in a job area in which his aptitude exceeds the level of his general ability. The Army has just that much more manpower at its disposal.

(2) An indication of the improvement in classification based on Aptitude Area scores is given by determining the number of men with above-average scores on AGCT and on their best Aptitude Area. Figure 15 shows the percentage of men with above-average

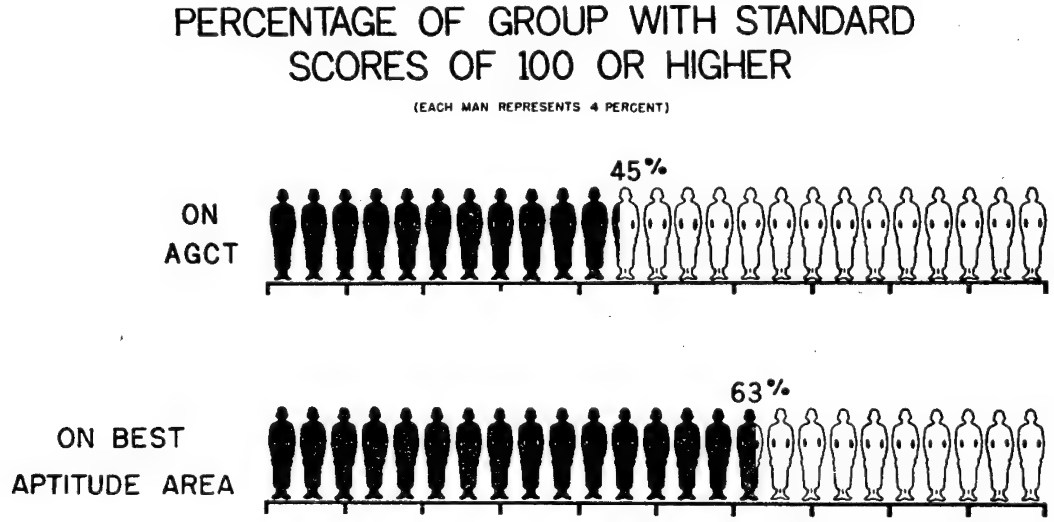


Figure 15. Percentage of group with standard scores of 100 or higher on AGCT and on best aptitude area.

aptitudes, as represented by standard scores of 100 or higher, among a sample of 1,132 enlisted men tested during one period at a training division. When these 1,132 men were tested with AGCT, only 45 percent had scores of 100 or higher. In contrast, 63 percent had scores of 100 or higher on their best Aptitude Area. Thus, the use of Aptitude Area scores permits a more effective classification of men at their higher skills and in more important Army jobs.

113. Problems in Classification by Aptitude Areas

a. Priority Assignments. Differential classification does not go on uncomplicated by the Army's manpower problems. For one thing, it is necessary to consider the relative importance of the various job areas. In order that certain important jobs may get enough high caliber men, it may be necessary to deflect whole groups of men to that area with little regard to their other Aptitude Area scores. Then, there are certain military jobs to which assignment is made on a first priority basis. That is, if a man meets minimum qualifications for the job, he is assigned to it, no matter what higher abilities he may have.

b. School Quotas. The filling of quotas for training in service schools presents an allied problem, since quotas are set on estimated future Army needs. The necessity for making the required distribution of incoming enlisted men to schools sometimes forces assignments which do violence to the principles underlying the Aptitude Areas.

c. Problems of Timing. If assignments could be equalized over a period of time, it would be possible to make a more consistent application of the Aptitude Area system. Unfortunately, it is not possible to control the distribution of abilities of men coming into a classification center at any one time. Nor is it always possible to anticipate the Army's needs which arise suddenly. And, often, the urgent need and the men qualified to fill it do not come at the same time. In these emergencies, about the only principle of differential classification that is almost surely met is that a man be not assigned in his poorest area. It can be said that with Aptitude Area scores available, the manpower supply is somewhat better catalogued than it was before. It is classified according to a finer breakdown of skills and abilities, which, in turn, permits a slightly greater adjustment to manpower requirements.

Section III. SUMMARY

114. Aptitude Measures as Aids in Classification

a. Initial classification attempts to make the best fit between the capacities of the enlisted men on one hand and the requirements of the jobs to be filled on the other.

b. The Army Classification Battery of objective tests was developed for use in evaluating the capacities of enlisted men for initial classification. With development of Aptitude Areas by combining two or more of the tests, and relating them to job areas, differential classification has become possible.

c. Differential classification involves matching the requirements of various job areas with the estimated strengths and weaknesses of men to be assigned. Other factors that must be considered are the relative importance of the job areas and other available personnel information.

CHAPTER 7

ACHIEVEMENT TESTS

Section I. CONSTRUCTION AND EVALUATION OF ACHIEVEMENT TESTS

115. What Achievement Tests Are

a. Tests whose purpose is to estimate in advance how well a man can learn a job are called aptitude tests. Tests whose purpose is to estimate how much a man knows about a job and how well he can perform the job are called achievement tests.

b. The various types of achievement tests used in the Army and their respective characteristics have been discussed in chapter 2 and only a brief summary will be presented here.

116. Types and Characteristics of Achievement Tests

a. Paper-and-Pencil Objective Tests. They permit economical coverage of large areas of knowledge; they are economical to administer and score in large numbers; the scoring is not subject to the scorer's varying judgments; they can be standardized.

b. Work Sample Performance Tests. Work sample performance tests are used where paper-and-pencil tests are not available or where it is essential to determine how well a man can manipulate tools and equipment instead of estimating it from the knowledge he shows on a paper-and-pencil test. They may be used as individual tests after previous screening by group-administered paper-and-pencil tests, and with illiterates and non-English speaking men. However, performance tests are not adapted for administration to large groups at a time; they are expensive to construct and administer and they are difficult to score objectively (par. 119).

c. Ratings as Achievement Tests. The usual achievement testing procedures may not yield satisfactory measures of acquired work habits and attitudes and skill in dealing with people. Rating methods of various types are customarily relied upon to measure how men compare with one another in these qualities. Ratings are deceptively simple—unless great care is used, they may be worthless. For use as measures of achievement, ratings should be in terms of concrete situations that are significant for specific purposes, rather than in terms of generalized or abstract qualities. Other requirements for effective rating procedures are discussed in chapter 10.

117. Requirements for Job Proficiency Measures

What is measured to determine proficiency for a job will depend upon an analysis of the job itself. The job may call for not only knowledge and skills but certain work habits and attitudes and certain personal qualities which may be equally important.

a. Possession of Pertinent Information. For any job and for any training program for that job, certain information is fundamental to adequate performance. In most military jobs, the man needs to know specific facts concerning the job to which he is assigned. But this is not enough. He also needs to know certain facts about jobs related to his own and about the functions and organization of the activity in which he works. Possession of pertinent information is not the only requirement of competence on the job, but unless the man possesses that, he is not likely to be as useful as other men.

b. Accessory Skills. As used in connection with Army achievement tests, skill refers to

practical effectiveness in performing a task. In defining the skills required in qualifying for a job, it is important to describe not only the action to be performed or the work product to be achieved, but also the typical situations in which the skill is to be displayed. It makes a great difference, for example, whether the clerk typist who is expected to turn out neat and accurate copy is to work from a typed rough draft with corrections noted, from a longhand manuscript, from shorthand dictation, or from a sound recording. While the action in using the typewriter is the same, the accessory skills are different and the rates at which the transcription can be done may be different.

c. Ability To Apply Knowledge and Skills. Aside from routine operations, practically every job in the Army requires some ability to solve problems. Changing situations, new materials, adaptation of tools and equipment, modified procedures, defective materials or tools—these are some of the conditions presenting problems which the soldier must solve by drawing upon his knowledge and skills. To measure this ability means to determine how well the men solve problems typical of those encountered in the normal course of duty. Making such measurements requires the development of standardized problem situations in which the men to be tested will have considerable freedom of choice in each critical stage. Both the development and administration of such achievement tests may be complex and time-consuming.

d. Possession of Suitable Work Habits and Attitudes. Personnel managers in industry have long recognized that a major factor in employee turnover, both discharges and resignations, is faulty attitude toward work or inability to get along with fellow employees. In the Army, the equivalent of employee turnover is frequency of transfers from one unit to another. Often men who ask for transfer, or who are recommended for transfer in advance of their normal periods of rotation in assignment, are those who fail to adjust to their duties. Examples of questions relating to suitable work habits and attitudes and ability to get along on a given job are—Does the man feel that the job he is doing is worthwhile? Does he take personal responsibility for observing the rules of the job, such as safety precautions and cleaning up his place of work? Does he have the initiative to learn and undertake work related to his assignment when there is need for it? Does he accept suggestions? Does he take over in emergencies? Does he supervise effectively? Can he work well in a team? The answers to such questions cannot be obtained from the usual kind of achievement test. Instead, it is necessary to depend on such techniques as ratings and interviews.

118. Planning Achievement Test Content

a. What is actually selected as test content will depend, as has been stated, upon a careful analysis of the job as well as on the suitability of the measuring instrument. Techniques for analyzing occupations, jobs, tasks, and work operations in terms of the functions performed are applied. These functions are then translated into qualities which the worker must possess in order to perform them satisfactorily and a preliminary estimate is of the relative significance of each quality.

b. Once it has been decided what the achievement test should measure and whether the need for the test is sufficiently great to justify the expense of construction the steps described in chapter 2 are followed. A test plan is prepared to make certain that there will be adequate coverage of the important aspects of the job. Items are constructed to sample these aspects reliably and to provide useful levels of difficulty. The test is then ready for field testing.

119. Are Paper-and-Pencil Tests Practical?

The statement is frequently heard that the work sample performance test is the more practical test—that is, that it bears the more apparent relationship to what is being tested—than the paper-and-pencil test. The criticism may be somewhat substantiated. However, the work sample test may not be as practical as it appears. For one thing, a relatively small sample of the job usually makes up the test, although what is examined on the test is examined intensively. In addition, as stated in paragraph 116, work sample tests are usually

harder and more expensive to construct and to administer efficiently on a large scale than are paper-and-pencil tests. Scoring is more difficult. The paper-and-pencil test, on the other hand, can be administered on a group basis fairly easily. It usually emphasizes a broad and extensive sample of the job content rather than an intensive one. Its practicality may not be apparent but, through field testing, its practicality can be determined and improved if necessary.

120. Evaluating Achievement Testing

The construction of achievement tests is not an end in itself. Such measuring instruments must be evaluated to determine their usefulness for the purposes for which they are developed. This chapter will consider the characteristics of objectivity, reliability, and validity as they apply to achievement tests.

121. Objectivity of Achievement Tests

a. For an achievement test to be objective it must be possible for anyone to administer and score the test and obtain the same score that any other scorer would obtain. Objectivity is important in obtaining equivalent scores from examinees who have done equivalent work on a given test, thus enabling comparisons to be made among examinees. To this end, methods of administration and scoring are rigidly prescribed. Objective items for achievement tests are constructed in such a way that there is, among persons who are supposed to know the correct answers, universal agreement as to what is actually the correct answer. Essay tests are relatively easy to construct and the temptation is strong to rely upon them. The correctness of the answers, even if a scoring guide is used, is not always easy to determine. Furthermore, scoring of essay answers is influenced by how well the examinee writes, and not by the correctness of his answer to the question asked. There is some mistrust of the notion of testing *recognition* of facts and details through objective items rather than *recall,* reproduction in writing, and the so-called "higher mental processes" of the essay types. This need not be—proper construction of test items can measure the more complex characteristics and provide objective scores. Whatever might be said of objective tests, there is abundant practical evidence that objectivity is basic to the construction of reliable and valid achievement tests.

b. In general, the more chance permitted for the exercise of judgment or discretion on the part of the scorer, the smaller chance there is for obtaining a reliable score. Since objective tests do not require the exercise of judgment in scoring, this is not a source of unreliability for them. One practical point should be kept in mind—large scale scoring would not be possible in the Army if objective tests were not used. To provide enough competent judges to score essay tests would require a small army of judges.

122. Reliability of Achievement Tests

To be reliable, an achievement test should—as should other types of measuring instruments used in personnel actions—yield about the same scores on repeated administrations of the same test or on alternate forms of the test. The techniques for determining and increasing reliability have been discussed in chapter 4.

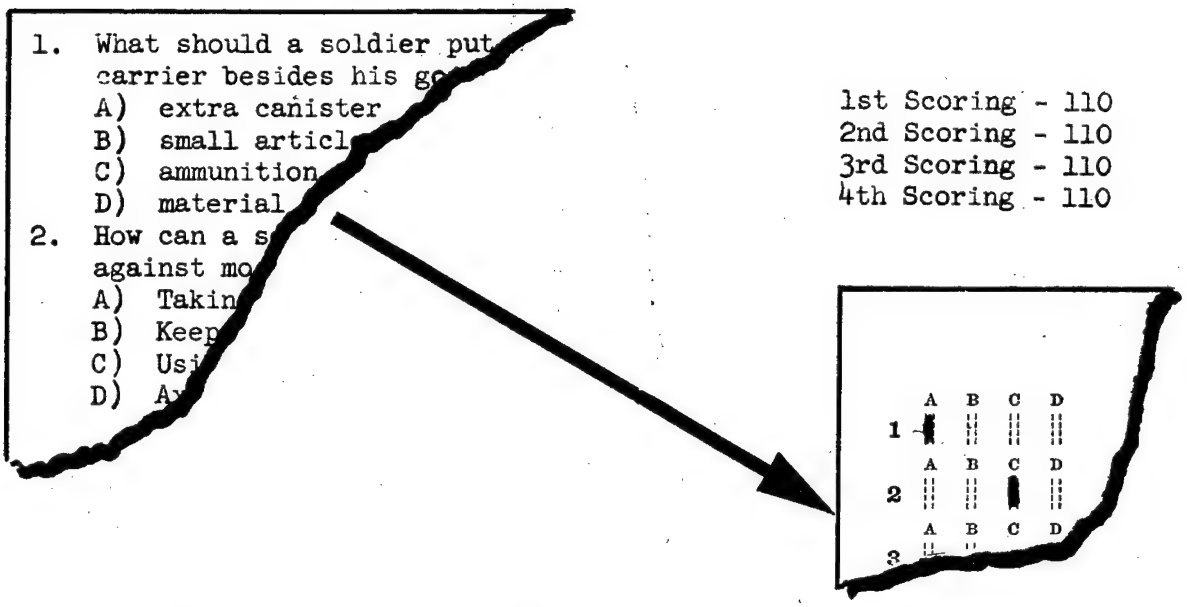
123. Validity of Achievement Tests

In testing achievement, the test constructor is called upon to select the achievements which are most likely to be significant for the purpose as specified, and to develop valid measures of these achievements. Validity of achievement tests may take various forms, and frequently the statistical demonstration is not complete. Some of the recent thinking on demonstrating validity of achievement tests is given below.

a. Test criteria are difficult to establish for many qualities in achievement measurement. Sometimes statistical or empirical validity is estimated for a new test by correlating scores on it with scores on other tests which experience has shown to be satisfactory. This procedure, however, is of uncertain value. The new and the old tests may be correlated, but the criteria they predict may be quite different.

b. The test constructor frequently attempts to build validity into his achievement tests. Constructing achievement tests is a long process and there frequently is not time to conduct

A STANDARD ANSWER TEST WILL GIVE A MAN THE SAME SCORE EVERY TIME SCORING IS REPEATED



BUT AN ESSAY TEST MAY GIVE HIM A DIFFERENT SCORE WHEN SCORING IS REPEATED

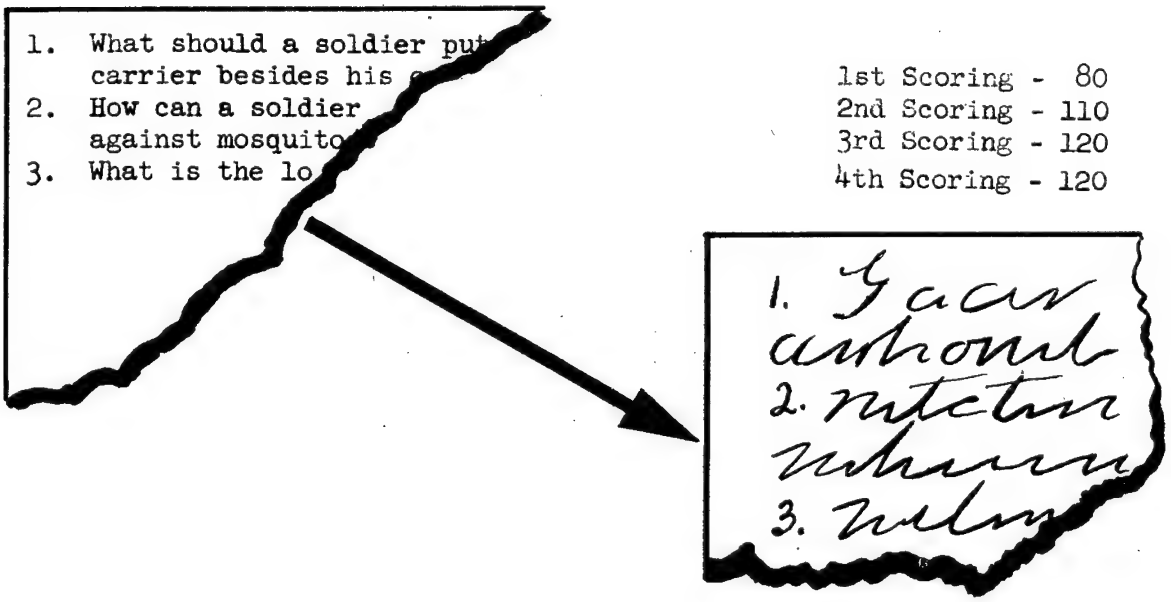


Figure 16. Comparative objectivity of standard answer tests and essay tests.

a field test to determine validity. For some jobs, there may not be enough men assigned in the occupation to provide an adequate population on which to validate the items prior to the official use of the test. The test constructor then relies on his seasoned judgment of what will constitute a "face-valid" test (ch. 5). This means the test must be designed and developed in such a way that there can be reasonable expectation that it will really measure the knowledge or skill which it is intended to measure. The test is a systematic sampling of the kinds of knowledge and skill the test constructor and experienced operating personnel consider crucial for differentiating among examinees. Much of the case for validity of such tests rests, then, not on a computed coefficient of correlation, but on an inferred correlation based on a description of job knowledge or training requirements, the representativeness of the items, and insight with which the items have been selected.

c. Constructing achievement tests may be likened to the development of many kinds of criteria (ch. 3). Subject matter experts provide the basic information on what is right or wrong, or good or bad. In the absence of empirical validity data, it is important that achievement tests have adequate reliability, a proper range of difficulty and adequate differentiation among men at various score levels.

d. When conditions permit, the achievement tests may be tried out in the field. The forms of the test are administered to a sample population to determine their range of difficulty and to determine their internal consistency (ch. 2). The results are analyzed and the test is revised, if necessary, to make certain that the difficulty of the test is suited to the population to be tested and that the test has sufficient reliability. With proper difficulty level and sufficient internal consistency, the built-in "face validity" may reasonably be expected to result in an empirically valid test.

e. In training courses of the classroom type, where there are usually not enough trainees available for a trial administration, it is not uncommon for classroom tests to be designed, developed, and issued for use without benefit of any experimental administration. Whatever validity the test may have is primarily dependent on the judgment exercised in choosing questions, problems, and situations which will make the test "face-valid."

f. Other factors relative to the construction and administration of achievement tests may have bearing on its validity. Tests of knowledge, for example, should not be tests of the ability to read. If too much material for the time allowed is included in a test measuring, for example, knowledge of light weapons operation, the test may become, in part, a test of how fast a man can answer questions. Or a test, by including very complicated items, may actually measure, in part, the examinee's ability to understand the complicated directions rather than what he knows about a given subject. Such tests may be completely invalid.

Section II. USES OF ACHIEVEMENT TESTS

124. Classification of Army Personnel

Screening and selection instruments are essential to the most effective assignment and promotion of personnel, and achievement tests can play an important part.

a. Screening is done when the administrative problem is to identify all those in the available supply of personnel who meet the minimum qualification requirements for a given type of duty. As mentioned in the first section, screening tests are frequently tests measuring the amount of knowledge and skill achieved up to that time.

b. Selection techniques may be used in the following situations:

(1) When the problem is to choose from among those available a required number who will be best qualified to perform a duty, or who are most likely to benefit from a training program, selections may be made from a screened group.

(2) In order to assign men who are partially qualified for a specific assignment or type of duty to an accelerated training program. Particularly dur-

ing mobilization, the services bring in large numbers of men who present varying levels of technical proficiency in civilian occupations. Many of these occupational skills are directly convertible to military jobs. Others can be used to a considerable degree in the operation and maintenance of military equipment that has no counterpart in civilian industry. Achievement testing may be used to help verify an individual's report that he has the knowledge and skills necessary to perform a given job adequately.

c. Promotion is a typical administrative problem in which both screening and selection measures are used. The Career Guidance Program offers an example of Army-wide use of achievement tests for purposes of promotion. Although use of tests for this purpose was temporarily suspended in 1950, they are still in wide use in making assignments and in estimating the effectiveness of training programs. A Career Guidance test is a test of proficiency in a military occupational specialty. Each such test was designed to measure the knowledge and skill a man had already acquired about the MOS for which he was seeking promotion. It was assumed that the greater an individual's knowledge about the work on a higher level the more likely he was to do the job ahead successfully. Needless to say, achievement tests were not used as the sole determiners of promotion.

125. Use of Achievement Tests in Training Programs

Achievement tests may be used for a variety of purposes in training programs. For one thing, training programs may not be accomplishing their objectives. It is necessary to administer tests to determine if the trainees are learning what they are supposed to learn.

a. Adjusting Instruction to Different Levels of Achievement. Achievement tests may be used to divide training classes into two or three groupings representing different levels of previous achievement. The instructor can then adapt the instruction to the level of the group.

b. Diagnostic Testing for Training. Sometimes coaching is necessary to provide special background knowledge and skill. For example, a trainee may demonstrate a high degree of background knowledge and skill necessary for an infantryman, but be very weak in the mathematical skills required to use maps effectively. In fact, an entire group may be lacking in this knowledge. Achievement tests may be used to discover specific areas of background knowledge and skill which require intensive teaching.

c. Measuring Progress in Training.

(1) A common use of achievement tests is to measure progress of trainees. If a well selected group shows a weakness on a test, the instructor may suspect that his instruction in that area has been inadequate. He may discover that the group understands a particular area and spend less time on it. And of course, he can discover the weaknesses and strengths of individual trainees.

(2) Not all subject-matter tests should be used as measures of progress or achievement. Sometimes such tests are primarily training aids—they need not be built according to the principles of good achievement testing if they are used primarily to arouse interest or serve as talking points. Short informal quizzes that make no attempt at adequate coverage may be used for this purpose. Generally speaking, they should not receive as much weight in determining achievement as do the larger periodic tests.

(3) As measures of progress, the larger periodic tests should serve to inform the trainee as to his progress and to motivate him to improve if necessary. This means that he should be informed of the results of the test as soon as possible—not only his total score but also his answers to individual questions.

d. Measuring Achievement at End of Course. Final examinations provide a basis for deciding which trainees have attained a satisfactory level of competence for the jobs for which the train-

ing was provided. The scores achieved also serve as an indication of how effective the course has been in training men to do the job. Accordingly, final examinations should be directed at the purpose of the training program, rather than at its content. They should show whether the man can use what he has learned in practical situations. Knowledge of facts, principles, or procedures may be essential to competent performance, but mere possession of knowledge is not a guarantee that the individual's performance will be competent.

e. Comparing Effectiveness of Various Training Methods.

(1) An achievement test may be used as a criterion to determine the relative effectiveness of various training methods. Merely trying out various methods is of little use unless an acceptable standard is used to evaluate them.

(2) Suppose, for example, it is desired to determine which of several methods is best for teaching recruits the nomenclature and functions of the various parts of an M-1 rifle. Various practical methods might be tried on different groups, such as formal lectures, demonstrations and informal discussion, manipulation of the weapon by the recruits, and so on. If care is exercised that the various groups of trainees are equivalent in background knowledge and aptitude, and if the various instructors are relatively equal in effectiveness, then a test administered at the end could show which methods are the more effective. An achievement test, then, can be used to evaluate the results of an experiment in training methods.

Section III. SUMMARY

126. Achievement Tests as Measures of Proficiency

a. Achievement tests are tests which reveal how proficient an individual has become as a result of experience or training with reference to a particular Army job. Most achievement tests take the form of paper-and-pencil tests or work sample tests. Achievement tests can measure not only information but skills, the ability to apply skills and knowledge, and the suitability of an individual's work habits and work attitudes.

b. In planning the content of achievement tests, attention is paid to economy through proper selection.

c. The chief requirements of achievement tests are objectivity, reliability, and validity. Empirical validity is frequently difficult to establish and recourse must be had to "building in" validity by selecting face valid items. It is consequently necessary that achievement tests have a proper range of difficulty and adequate differentiation at various score levels.

d. Achievement tests have been used chiefly as aids in classifying and promoting Army personnel and as aids to training. In the service schools, achievement tests are used in estimating backgrounds of trainees, and in measuring their progress and their achievement at the end of the course. Achievement tests may also be used to measure the effectiveness of various training methods.

CHAPTER 8
INTERVIEWING AS MEASUREMENT

Section I. PURPOSE OF INTERVIEWS

127. General

The interview is one of the oldest personnel techniques and one of the most frequently employed in the Army. It is a technique in which a great many people feel they are competent and which they do not hesitate to use.

128. Fact-Finding Interviews

Various types of interviews are employed in classification of enlisted personnel. The requirements for conducting these interviews are described in the special regulations on classification procedures for enlisted personnel. * Essentially fact-finding interviews provide the interviewer with a basis for classification or to provide guidance. In general, interviews should not be used for fact-finding if other more economical methods of obtaining information are possible. Frequently, however, other sources of information are not available or are inadequate; the interview may then be used. Where these sources are adequate, it is wasteful to use the interview.

129. Measurement Interviews

Interviews may be used as measuring instruments to yield objective scores. When so used, they are constructed, standardized, and validated in a fashion similar to other personnel measurement techniques. Such interviews may have different purposes which are reflected in their content and in the way they are conducted.

130. Administration of the Measurement Interview

If interviews are to be applied in personnel measurement, they must be used and scored exactly in accordance with the particular manual for the conduct of the interview which governs, if they are to yield results comparable to those obtained when they were standardized.

Section II. THE INTERVIEW AS A MEASURING INSTRUMENT

131. Limitations of Interviews as Measuring Instruments

The measurement interview, when conducted by qualified personnel strictly in accordance with the appropriate manual, will yield results which make a useful contribution to problems of personnel evaluation and selection. One such interview, structured and conducted to assess specific aspects of the interviewee, is used in the selection of candidates for Officer Candidate School. However, existing evidence suggests that such interviews are quite sensitive to the interviewers' deviations in procedure from that prescribed in the manual. When such deviations occur, the resulting scores very commonly have little reliability and no validity. Even under the most favorable conditions, the validity unique to the measurement interview tends to be small.

132. Practical Considerations Governing Use of the Interview

As a personnel measuring instrument, the interview tends to be a most expensive one. To obtain reasonable reliability of scores, at least three interviewers are usually required. Also,

* SR 615–25–25, 7 February 1951, Enlisted Personnel, Classification Procedures (sec. IV).

only two to four interviewees can be processed in an hour, with a probable maximum of from 20 to 25 per day. On this account, and because of its sensitivity to even slight deviations from the conditions prescribed, the measurement interview is not a technique to be used indiscriminately. Its use is justified only where it can be employed precisely as prescribed, and when no other means of obtaining the desired result seems practicable. Furthermore, it should not be used as the sole instrument in a selection program.

Section III. SUMMARY

133. Interviews as Personnel Measuring Instruments

a. Interviews have not, in general, proved valuable in estimating abilities. They may be used profitably in guidance. They are also useful in obtaining information, although the same information may sometimes be obtained more economically by other means.

b. The standardized measurement interview has some usefulness in estimating certain aspects of the individual. It must be conducted according to the prescribed procedure. Practical considerations limit its use.

CHAPTER 9

SELF-DESCRIPTION TECHNIQUES

Section I. THE NATURE OF SELF-DESCRIPTION QUESTIONNAIRES

134. General

Information about how well a soldier may be expected to perform can be gathered from many sources which are outside the man himself (efficiency reports, special evaluations, school grades, and interviews) and are described in other chapters. Aptitude and achievement test results, where the information comes from the person himself, also have been described. In this chapter will be discussed the person's self-descriptions, his self-evaluations, his own statements regarding his background, his attitudes, beliefs, and personal reactions. The typical method of gathering such information is by "self-description techniques" in the form of biographical information blanks and personality questionnaires. The method of developing such forms, and their use in selection and classification, is described in this chapter.

135. Kinds of Self-Description Data

There are two kinds of self-description data. First, there is the objective, factual background information. Such information as how far a man went in school, what kind of work he has done, his family's social and economic status, etc., fall into this group. The second kind of information is much more subjective. It concerns primarily what the man thinks about himself, how he feels about his environment, also his attitudes and beliefs regarding other people. Both of these types of information may be gathered in a single form, which might be called a "self-description form," a "biographical information form," a "personality inventory," a "preference form," or by some similar name. In any event, the title itself is not necessarily a guide to the content of the self-description instrument.

136. Advantages of Self-Description Questionnaires

What are the advantages of self-description blanks, as compared with other means of gathering information about a person? When and how may they be used to best advantage? These questions will be discussed in the sections that follow.

a. They provide a fairly simple and systematic means of gathering information, supplied by the person, about himself in a manner which permits its scientific evaluation. From the scientific standpoint, little is known about the significance of the various bits of information about a person which might be gathered from the person himself or from other sources. The mass of data that can be gathered is enormous. Its very volume presents a serious problem of organization. However gathered, before interpretations can validly be made, it is necessary to make a careful analysis of the data. If the information is to be used for measurement purposes, it is necessary to know which items of information are really useful in prediction. After this is discovered, a systematic means of evaluating and using the information about each individual is needed. Self-description forms have been so devised that the information gathered can be readily and systematically studied and utilized.

b. Self-description forms provide standard conditions. The wording of questions is very important in determining the responses obtained. It is practically impossible to maintain uniformity on such things in an interview. Even if the wording of the question is not varied, such things as inflection, rate of speech, and emphasis may influence the interpretation of

a question. The individual asking the question tends, often unconsciously, to put his own interpretations and reactions into the question. Furthermore, the personal nature of some of the questions may make many people hesitate to reveal their true feelings in an interview. The printed instrument, by contrast, is impersonal. In appearance it is uniform, always exactly the same for each person. However, special directions may be necessary so that all persons filling out the blank adopt, as nearly as possible, the same attitude toward the form. Even with these precautions, there is strong evidence that important variations remain in the attitudes of individuals responding to the questions, and that these variations have an effect upon the results obtained. Some of the methods employed for controlling the effects of these differences in attitude will be described later in this chapter. These methods aid in making self-description instruments generally more valid and fairer to the men to whom they are applied.

137. Uses

At present, the most frequent use of self-description forms is as an aid in screening applicants for certain special assignments where the more intangible personality traits appear to be important. Chief among these are leadership positions. Self-description forms have been developed for OCS, for Leaders' School, and for selection of ROTC honor graduates for commission into the Regular Army. Self-description forms for selection for a number of enlisted specialties have also been devised, and the number is growing constantly as further research is accomplished.

Section II. CONSTRUCTING THE SELF-DESCRIPTION FORM

138. General

Self-description forms are more than a series of questions, as already suggested. It is important to understand the difference between an Army self-description form and the typical questionnaire. To point out this difference some questions may be asked. Just how is such a form constructed? What steps are necessary to obtain usable results? How can we tell how good a self-description form is after it is constructed? The following sections give the answers to these questions and lead into other questions regarding means of making these instruments as valid as possible.

139. Constructing Items for Self-Description Form

To begin with, it is necessary to invent or borrow a large pool of self-description items which are intended to be related to whatever one is trying to predict. Let us suppose the problem is one involving the prediction of leadership. What is important in leadership? What traits of character, what home, family, school, and employment backgrounds may possibly contribute to success in leadership? What personality quirks are likely to prevent a person from being a good leader? What personal habits and mannerisms, what patterns of likes and dislikes, what beliefs and attitudes appear to be important? No one knows, for sure, before try-out, whether any particular trait is important or not. No one knows, before try-out, which items are related to some important trait. Experience over a number of years indicates that only a small percentage of self-description items written are actually found to have the necessary validity. In view of this, many more items are written than are needed for any particular form. Examples of items which might appear on a self-description form are shown in figure 17.

140. Preliminary Validation of Experimental Self-Description Items

a. The Criterion. After a large number of items have been written, from five to ten times as many as one ultimately expects to use, the items are tried out to determine their validity. The first step, as in the field testing of other instruments, is the development of a criterion. The considerations involved in this process, the most difficult task in developing instruments, have been described in chapter 3. To validate self-description items for prediction of leadership ability, it may be decided that anonymous ratings by close associates and by immediate

SELF-DESCRIPTION ITEMS

CAN REFER TO FACTUAL BACKGROUND

Check one

How many living brothers and sisters have you?

A) none
B) 1
C) 2
D) 3
E) 4 or more

When of high school age, how many evenings a week did you go out?

A) less than 1
B) 1
C) 2
D) 3
E) 4 or more

PERSONALITY CHARACTERISTICS

Mark A or B

A) I believe it is important to get what you want even if you have to fight to get it.

B) I believe it is more important to be well-liked by my employees than to get the work done exactly according to established procedures.

A) I am open-minded.

B) I hold my purpose.

Mark the degree to which the statement applies to you.

Always finish what I start

A) To a low degree.
B) To the usual degree.
C) To a high degree.

Figure 17. Types of items used in self-description forms.

superiors provide a satisfactory index of leadership effectiveness. The special procedures for obtaining these ratings are described in chapter 3. These criterion ratings are obtained for the group of men who answer the questions in the self-description blank.

b. Determination of the Validity. Once the criterion ratings have been obtained, the items are tested against the criterion. This process consists, in essence, of finding what percentage of men who received high criterion ratings answered "yes" or "high degree" to an item, and comparing it with the percentage of men who received low criterion scores and who answered "no" or "low degree" to that item—in other words, how well the item "predicts" the criterion ratings. An item responded to one way by men rated high and the opposite way by men rated low would be considered valid and retained. If both good and poor men gave the same kind of answer, the item would not be considered valid for differentiating between them. This process of item analysis is a laborious one, but it is essentially the same process as described in chapter 2 in connection with the construction of any test. With self-description instruments, however, it is essential. There is no way of knowing whether a self-descriptive item is valid or not until it is tried out. It may be a desirable quality for an item to appear to be valid, but this apparent validity cannot be counted on in any degree as a substitute for demonstrated validity. As mentioned before, only a minority of items can be expected to have sufficiently high demonstrated validity to warrant their use.

c. Selection of Items. Once the validity of each item has been determined, the instrument can be set up in form for actual use. The items which have proved to be valid are either assembled in a new form or are spotted in the original one. A scoring key is set up to permit counting the number of the valid answers chosen by each person who takes the test. The score obtained from such a key is the revised predictor of the criterion. The problem remains of determining how accurate a prediction it can give.

d. Cross-Validation. To find out how good the self-description form really is, it is necessary to try the revised form out on a new group of men on whom criterion ratings are obtained. If good agreement is found between the criterion ratings and the self-description answers as scored by the key developed from the item analysis, that is, if the key predicts the criterion for this new group, there is reasonable assurance that there is a real relationship between the keyed items and the criterion. This process of second try-out is known as cross-validation (par. 32). If this second try-out fails to show substantial relationship, it is necessary to discard the key, and start over again with either another group of items or a more stable criterion, or both. This important step of cross-validation is one which has been emphasized in developing Army measuring instruments to assure their validity.

Section III. SUPPRESSOR METHODS OF IMPROVING VALIDITY OF SELF-DESCRIPTION FORMS

141. General

Self-description instruments are subject to a number of sources of error due either to conscious or unconscious biases. That is, people give answers which they think will create a favorable impression or will help them to get a desired assignment rather than the answers that are strictly true. Control of these sources of error has been a major problem and as yet it is far from solved. Two general methods have been used—the suppressor method and the forced-choice method. These methods and the degree of success achieved with them will be discussed next.

142. Selecting Responses to Yield a Desired Score

Bias in the results of self-description forms may occur in two ways. In one way, the person filling out the form selects the face-valid items; that is, he marks those items which he thinks will give the score that he desires. In the other, (par. 143), the responses he selects are deviations from accurate statements and distort the self-appraisal.

a. The Face-Valid Item. A man may consciously try to obtain as high a score as possible. He will then choose answers which he thinks will give him a high score. If, for ex-

ample, it is intended to predict leadership effectiveness, then some items "on their face" will appear to be related to leadership. The most obvious example of such an item would be "I am a good leader"—to be answered yes or no. Other items might not appear to be related to the criterion at all. It is possible to get an estimate of the extent to which each item appears to the average person filling out the form to be related to the leadership criterion. The men to whom the form will be applied can be asked to rate the items on a scale, indicating how important they feel a "yes" response would be in indicating high leadership potential. The average of these ratings for any item would then be its "face validity" index. Various means can be used to apply this knowledge to improve the validity of items. One is the development of a "suppressor key." This name comes from the fact that such a scoring key will tend to allow for or in other words, act to reduce or suppress the effect of a person's choosing items on the basis of their face validity.

b. Adjusting for Face Validity. If every one reacted in exactly the same way to face validity, there would be no need to allow for it. The men would all keep the same relative positions in the group. All scores simply would be increased by a fixed amount. But men do not all react alike to face validity. Some will change their responses little, if at all, in their effort to try to obtain a high score. Others will be influenced greatly. Their responses may indicate more about what they think they should say than what is actually true of them or what they really think of themselves. Most people will fall somewhere between these extremes. The important point is that people differ greatly in this respect. If our measurements are to be as valid as possible, we must measure and allow for this kind of difference between individuals. This may be done by including in the test a number of items with high face validity but no real validity when compared to the criterion. Persons who choose a large number of such items are reacting on the basis of face validity, that is, they are responding on the basis of what they think will give them the high scores or low scores they desire, rather than marking the responses that are really characteristic of them. A key ("suppressor key") is then set up to score these face valid items. If the score on such a key is substracted from the score on the valid key, a corrected score is obtained which makes allowances for the effect of the face validity of the items.

143. Distortion as Source of Error

a. The Tendency to Over-Estimate or Under-Estimate. Important as face validity is as a source of error, it appears that another source of error, called distortion, is even more important. Such distortion may be quite unintentional and even unconscious. It is possible that it reflects a person's deep-seated attitude toward himself. A conscientious effort to be "honest" in self-evaluation is not enough to deal with some of these attitudes. Some people are just naturally over-optimistic about themselves, while others are unduly modest. A strong warning to be "honest and objective" in answering a personality questionnaire may have little effect on the first kind. They just know they're good and say so. The second kind, on the other hand, may lean over backwards in describing their merits. They may be far more frank than average in revealing their shortcomings. A means must be found to allow for such exaggerations which distort self-appraisal. The measurement of this distortion is a much more subtle and difficult task than measurement of face validity. The solutions to the problem that have been found so far are rather complex but serve to indicate the importance of painstaking detail in constructing valid self-description forms.

b. The Distortion Score. Even on a self-description form that predicts a particular criterion fairly well, distortion will cause some men to get unduly high scores, other men to get low ones. From the standings of the men on the criterion, what their self-description scores *should be* can be determined if the degree of relationship between the instrument and the criterion has been previously determined. For each person, the discrepancy can found between the score he *should have* and the one he actually obtains when he fills out a self-description form. This discrepancy may be referred to as *distortion score.* It is a measure of the tendency of

each person to distort his responses, so that they are either unduly high or unduly low.

c. Allowing for Distortion. The next step is to find out which items are a measure of this distortion tendency. To do this, an item analysis is made to discover which items are responded to in one way by the men with high distortion scores and responded to in another way by those with low distortion scores. When these items have been identified, a suppressor key is set up to allow for the effects of distortion. A person who is over-optimistic in his views of himself has, in effect, a subtraction from his score. A person who is leaning over backwards, on the other hand, has an increment added to his score. This process seems to have considerable promise in increasing the validity of self-description instruments.

Section IV. FORCED-CHOICE METHOD OF IMPROVING VALIDITY

144. General

Thus far two sources of bias or systematic error in self-description blanks have been described and means of dealing with them using the suppressor method have been discussed. There may well be other methods, and research is continuing on methods of improving validity. One of these methods is known as "forced-choice." See chapter 10 for application of this method to rating procedures.

145. What is a Forced-Choice Instrument?

In a typical forced-choice instrument, the person is required to respond to each item by choosing between two alternatives which appear equally attractive, but only one of which is valid. There might be three alternatives, and the person may be asked to choose the one which least describes him and the one which best describes him. Or there might be a list of a dozen phrases, and the person required to choose the five that best describe him, and the five that least describe him. However many alternatives are presented, they are so set up that the valid alternatives appear as attractive as the non-valid alternatives. The attractiveness (par. 146) and the validity have been previously determined.

146. Grouping Alternatives

For the sake of simplicity, only the grouping of forced-choice alternatives into pairs will be described. The procedures employed apply to larger groupings as well. The basic procedure for grouping alternatives takes into account the two main characteristics of each alternative —its attractiveness and its validity.

a. Three measures of attractiveness are obtained. (1) The number of times an alternative has been marked by the men as describing themselves (the "p value"), which might be called its popularity and which is similar to the difficulty value of test items (ch. 2); (2) the face validity index, (par. 142); and (3) the distortion index, (par. 143).

b. In principle, the alternatives which would be paired would have the same attractiveness, as defined above, but different validities, one alternative having high positive validity (the alternative chosen only by competent men) and the other having zero or negative validity (the alternative chosen equally often by competent and incompetent men or chosen only by the incompetent). In other words, the man should have considerable difficulty in choosing one of the alternatives except on the basis of whether it really describes him.

c. In practice, as usual, compromises are necessary. The alternatives cannot always be paired on the basis of all three attractiveness measures because of inconsistencies in the data. Thus far it has been possible to use "p value" and the distortion index together or the face validity index and the distortion index. Other compromises include using alternatives whose attractiveness measures are not as alike as would be desired, or those where the differences in validity are not as great as would be desired.

d. The grouping of alternatives as just described is not applicable to all kinds of material. Factual background items do not lend themselves to forced-choice treatment. It would not make sense to ask a person, for instance, which is more true of him: (a) that he had an eighth grade education, or (b) that his father was American-born. These things are true or not

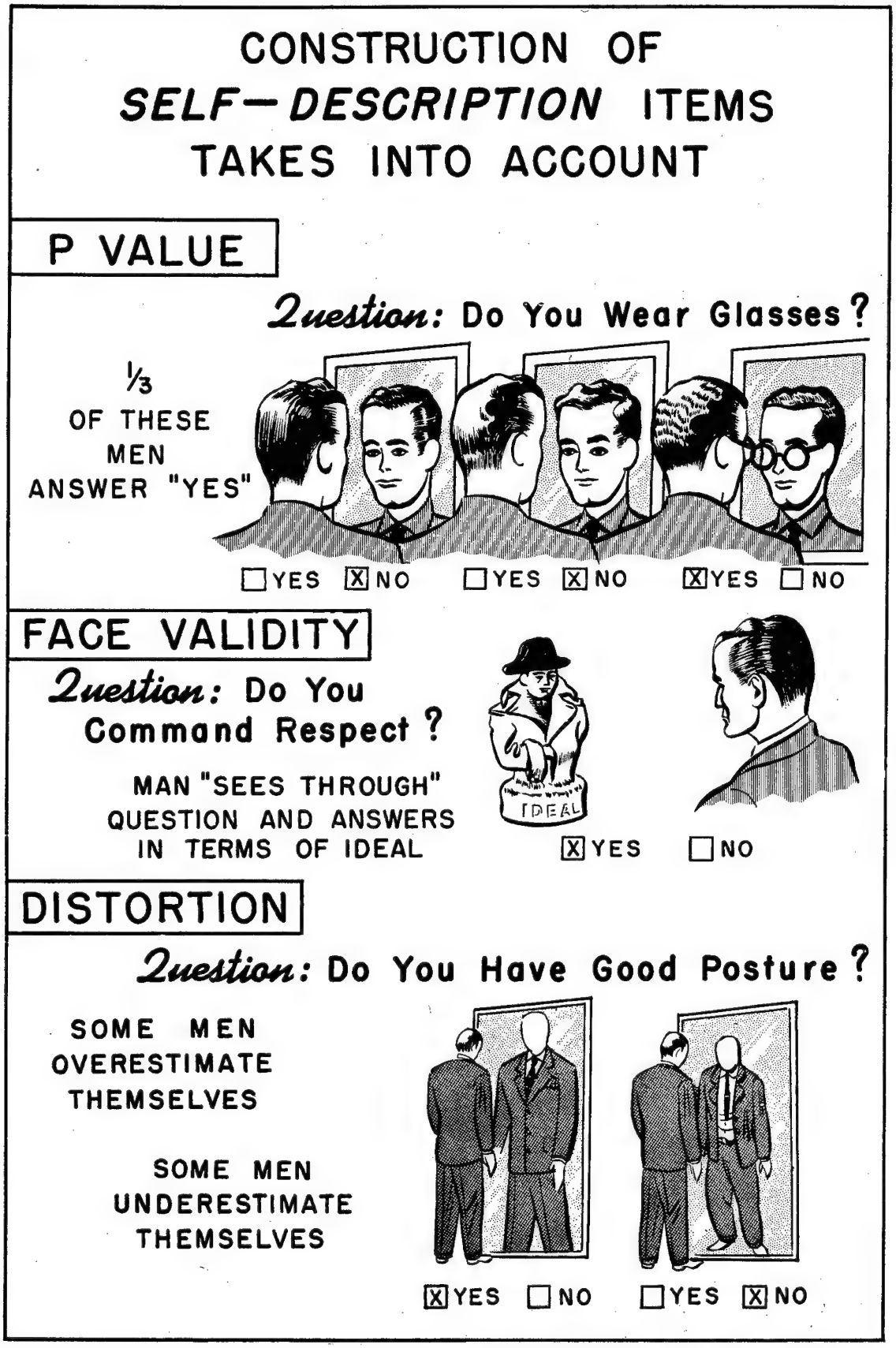


Figure 18. Important characteristics of items considered in developing self-description forms.

true, and no weighing of evidence or judgment is involved in answering them. There is, of course, no hard-and-fast line between the purely factual and purely subjective items, where forced-choice may be applied. One type grades off into the other. Even within the area which may reasonably be considered purely subjective, there are logical limits to the freedom with which items may be paired. It would not be reasonable to ask a person, for instance, which he liked better, horseback riding or tomato soup. In spite of these limitations imposed by the meaning of the terms and their statistical characteristics, it has been found so far that the forced-choice procedure applied to self-description forms produces higher validities than does any other method.

147. Checking Validity of Scoring Key

The construction of a number of forced-choice items on the basis described above does not complete the task of building a forced-choice self-description form. Responses to items in pairs are not always the same as to each member of the pair presented by itself. Thus, there is no assurance that an alternative found to be valid when it was not paired will maintain its validity when paired with another alternative. A key scoring the valid alternatives which are selected can be prepared on the basis of the individual alternatives in each of the items. Such a key is known as a "predetermined" key. However, there is no assurance that this key will stand up after the pairing. Before the forced-choice self-description form can be said to be ready for operational use, it is necessary to check the validity of the predetermined key.

a. To do a thorough job of building a forced-choice self-description form, the experimental instrument should be administered to two new groups of men for whom criterion data are available. The first of these groups is used for re-analyzing the validity of the items, that is, for constructing a scoring key based on the items after pairing. This is known as an "empirically determined" key as distinguished from a "predetermined" key.

b. After the empirical key has been set up, it is necessary to determine its validity on a second group (cross-validation).

c. Extensive research has shown that if the original samples are large enough and the criterion adequate, the empirically determined key is only slightly more valid than the predetermined key. For this reason the item-analysis after pairing is sometimes not performed. In that case, the experimental forced-choice form is administered to only one population. It is the predetermined key which is cross-validated. If the validity is found to be satisfactory, the form is then ready for operational use.

Section V. SUMMARY

148. The Value of Self-Description Forms

a. Self-description forms, under a variety of names, are used as a ready means of gathering, analyzing, and interpreting in a standard manner a vast and varied mass of facts which soldiers supply regarding themselves.

b. Successive steps in building such self-description forms are—

(1) Deciding on traits which seem to be related to the criterion.
(2) Writing a large number of items which appear to be related to the criterion.
(3) Experimental try-out and item analysis.
(4) Cross-validation.

c. A number of technical refinements have proved valuable in reducing biases. These methods include measurement and allowance for the popularity of the items, their face validity, and their susceptibility to conscious or unconscious distortion.

d. One means of increasing validity of self-description forms is the use of the forced-choice technique. This procedure requires a person to choose between two or more apparently equally attractive alternatives, one of which is more valid than the other. Where forced-choice procedure is not practicable, special keys are set up to measure and compensate for such sources of error in prediction as response to face validity and susceptibility to distortion.

e. Research now under way may extend the usefulness of self-description instruments from their present applications, chiefly in estimation of leadership potential, to general classification purposes.

CHAPTER 10

RATINGS AS MEASURES OF USEFULNESS

Section I. GENERAL CHARACTERISTICS OF ADMINISTRATIVE RATINGS

149. How Ratings Differ from Tests

a. Need for a Direct Measure of Usefulness.

(1) When an achievement test is administered to a man, what is learned is how much he knows or how well he can do. When an aptitude test is used, what is learned is how well he can be expected to learn. When a standard interview or a self-description form is used, what is learned is how well he is motivated, how well he can lead others, or perhaps how suitable his personality characteristics are for a particular job. The more valid the instrument and the more adequate the criterion, the greater the safety with which the information provided by the instrument can be used.

(2) There is still another kind of information needed for effective utilization of manpower. How effective is the man on his present job? How valuable? How useful to the Service? Not just useful as a map reader, or a marksman, or a first-aid man, but useful as an all-round infantryman. Furthermore, what does his present usefulness as an infantryman indicate about his future usefulness as a platoon sergeant, an officer candidate, a regimental personnel officer, a company supply officer?

(3) It is theoretically possible to administer a large number of achievement tests, aptitude tests, standard interviews, self-description forms, and other such instruments and obtain an average score which would indicate over-all usefulness. However, these instruments, although valuable, would be indirect measures. What is needed is a more direct measure, a measure based on a man's present job performance as judged by those who know the man's work. A comprehensive set of tests might give this information up to a point but it would not indicate what the man's superiors or associates thought of him or how they judged his worth.

(4) It is for this purpose that various rating procedures are used—to provide information on how valuable a man is judged to be by those who are in a position to judge his worth.

b. Additional Problems Involved in Ratings. The prime characteristic desired in any personnel measuring device is validity—is it useful for a specified purpose? Next in order comes the problem of reliability—is it a consistent measure? Tests of all kinds can be effectively dealt with in terms of these two concepts. Ratings need to be evaluated in terms of these concepts, but with ratings, these concepts are complicated by a number of conditions which are of relatively little concern when dealing with tests. Most of the differences stem from the fact that a rating involves not only a ratee, but a rater. A person may or may not like a test, but his performance on it is his own. His rating, however, depends not only on what he does but on how the rater values what the ratee does and how this evaluation is reported. The rater has definite ideas of where, on a given scale, he is placing the ratee and he expects administrative action to be consistent with this. The various

systems of values involved make the problem of acceptability of a rating procedure to the rater of extreme importance. The problem of validity and reliability are thus cast in a different setting, and a whole set of problems concerned with the characteristics of the rater and the conditions surrounding the rating are added.

Section II. PURPOSES OF RATINGS

150. Ratings Classified According to Purpose

The purposes of ratings may be classified under three main headings—guidance, administrative uses, and as criteria. This chapter is primarily concerned with ratings for administrative purposes. Confusion of these purposes is prevalent enough that a brief discussion to clarify the issues will be undertaken. One confusion is especially unfortunate—attempting to use the same procedure for evaluating a person for administrative actions (promotions, selection for assignments of unusual responsibility) and for guidance purposes (assisting a subordinate to improve in his work).

a. Ratings for Guidance Purposes.

(1) A prime responsibility of an officer or a supervisor is the development and improvement of the work of his subordinates. If a subordinate is to be helped, his strengths and weaknesses need to be analyzed in relation to the demands of a particular job. When this is done, the result will probably be a list of specific requirements, so specific, in fact, that they may sometimes apply only to a single individual. A rating scheme to serve this purpose must, therefore, be flexible and not be limited to a few characteristics or general job requirements. It does not lend itself to scoring. Even if it did, the score would be of little value in aiding subordinates to correct their weaknesses. At most, such a score could indicate only the general level of progress.

(2) Scoring this type of report is not only unnecessary, but undesirable. Tendencies to rate on general impressions rather than on specific aspects of behavior, to be lenient, and to render ratings with little spread are more likely to operate when ratings are to be scored. This defeats the guidance purpose and may well make the evaluation of uncertain usefulness.

(3) A rating procedure for guidance purposes requires an interview with subordinates. Otherwise, the subordinate will not learn of the results of his supervisor's analysis of his strengths and weaknesses. The guidance purpose would then be defeated.

b. Ratings for Administrative Purposes. The problems of obtaining a rating on job performance or over-all worth are different from those involved in rating for guidance purposes. First, ratings for administrative purposes should reflect a man's standing in his competitive group. A quantitative score is required. Second, their basic purpose involves evaluation of a person. Third, ratings for administrative purposes count; they influence decisions regarding a subordinate's career. This third characteristic has many implications, as will become clear. Fourth, administrative ratings are predictors, and, as such, require consideration from the standpoint of reliability and validity, exactly as any other predictor instrument; their value cannot be accepted, but must be established. Fifth, in such ratings the problems revolving around attitudes of the rater (acceptability of the procedure, for example) are more important. One example will suffice—the relationship between superior and subordinate involved in acceptance or nonacceptance of a rating procedure, and the effect of this relationship on morale. Sixth, ratings for administrative purposes do not require an interview between the rater and ratee. It does not prevent one, although the greater leniency that accrues under such circumstances tends to reduce the value of the rating. In general, ratings for official administrative purposes involve more problems than is the case of ratings for the other two purposes.

c. Ratings for Criterion Purposes. While a good many problems concerning leniency, halo, and other human tendencies which have been established with respect to ratings for both criterion (ch. 3) and administrative purposes are similar, the differences that exist make it unsafe to generalize from findings concerning one to the other. Both types of ratings require a quantitative score to reflect competitive positions in some group. Criterion ratings are not used for administrative purposes; their basic purpose is to evaluate a test or procedure, not a person. Criterion ratings are not predictors; they are accepted as the best index or yardstick of a particular kind of success that is available. Their value is accepted and validity studies are not conducted. In general, the differences between administrative and criterion ratings stem from the difference in the basic purpose—the one, to evaluate a person for administrative purposes; the other, to evaluate a test or procedure.

d. Specificity of Purpose and Effectiveness in Rating. Within the three major categories described above, rating purposes may be still more specific. Ratings for administrative purposes may be aimed at promotion at critical levels in the supervisory scale, such as promotion from company to field grade and from field grade to general officer. Or they may be aimed at selecting men for particular kinds of assignments, such as staff or line duty. In general, the more specific the purpose can be, the greater the likelihood that a better job can be done with a rating scale. This possibility of greater effectiveness must, of course, be balanced against the administrative problems involved in having a number of different scales, each for a specific administrative purpose. Thus far in this chapter, the word "rating" has been used. To keep clear the major purpose of this chapter, which is to discuss the problems and procedures appropriate to administrative ratings, such ratings will hereafter be referred to as efficiency reports.

Section III. EFFICIENCY REPORTING METHODS

151. Need for Efficiency Reporting Arises from Size of Organization

In a small business where all employees are under the immediate supervision of the owner, the problem of efficiency reporting does not arise. The owner has ample opportunity to observe all his employees. When the time comes to promote someone, he has little trouble in establishing an order of merit. As his business grows and becomes departmentalized, his problem of locating the best employees becomes increasingly difficult. And when he begins to have branch offices scattered all over the world, devising a technique of rating that will be valid for his purposes and fair to the subordinates becomes very difficult indeed. No less difficult is the Army's problem of rating fairly and effectively its great numbers of officers and enlisted men. The necessity for evaluating the usefulness and potentiality of subordinates makes the search for good methods a highly urgent one. Not only is the task urgent from the standpoint of making the best decisions on promotion, assignment, and other personnel actions, but it is equally urgent for its implications for the morale of the individual. Methods that reduce such differences as that between "hard" and "easy" raters are needed to increase the fairness with which efficiency reporting systems operate.

152. Kinds of Efficiency Reports

Previous attempts to improve efficiency reports utilized the methods employed in rendering criterion ratings. Many were found not suitable for use under administrative conditions. One, the forced-choice procedure, was applied first to the administrative activities and is just now being considered for use in criterion measures (ch. 3).

153. Rating Methods which are not Suitable for Efficiency Reporting

a. Essays—Descriptions and Evaluations of the Ratee in the Rater's Own Words. Essays have the advantage of permitting the rater to express himself as he feels. To offset this ad-

vantage are at least three serious disadvantages. One is that essay evaluations do not lend themselves to objective scoring; hence, they tend to have low reliability. Although it is possible for a group of judges to evaluate the essay evaluations and arrive at a score, this procedure is far too time-consuming for normal use. The second serious deficiency in essays is that the various raters are not likely to cover the same aspects of the job or the same characteristics of the man, thus making comparison of men difficult. The third disadvantage is the bias introduced into the evaluations by differences in the literary skill shown by the raters. In general, essays are not suitable for large-scale rating operations.

b. Ranking—Placing Men in Order of Merit. Ranking forces the rater to consider all men in relation to each other. He cannot be lenient and evaluate all his men high—he must discriminate among his men. Differences in rater standards are eliminated. However, ranking is not practical for general use. It cannot be used when only one man is being evaluated. The numerical score assigned each man depends on the size of the group, thus distorting the evaluations; for example, the worst man in a group of five would receive a rank of five, the same as the fifth man in a group of twenty. Comparisons of men from different groups is difficult without converting all ranks to a common scale. Finally, ranking shows relative order; it does not show the amount of difference. For example, in one group, men ranked 1 and 2 may be nearly alike, but in another group, men ranked 1 and 2 may be very different. Thus, rankings are seldom suitable for administrative use. For similar reasons, that variant of ranking, the nomination method, is also unsuitable.

c. Paired-Comparison Method — Comparing Men, Two at a Time. This method (ch. 3) has the same disadvantages as ranking. It can be applied only in very limited conditions. It is impractical in the Army for large scale administrative use.

d. Guided Ratings. Assistance of the expert in obtaining ratings, described in the discussion of criterion rating as a method of educating the rater (ch. 3), is a time-consuming and expensive process. It has promise in collection of criterion data, where the time and expense is justified. However, this method is impractical for efficiency reporting in the Army.

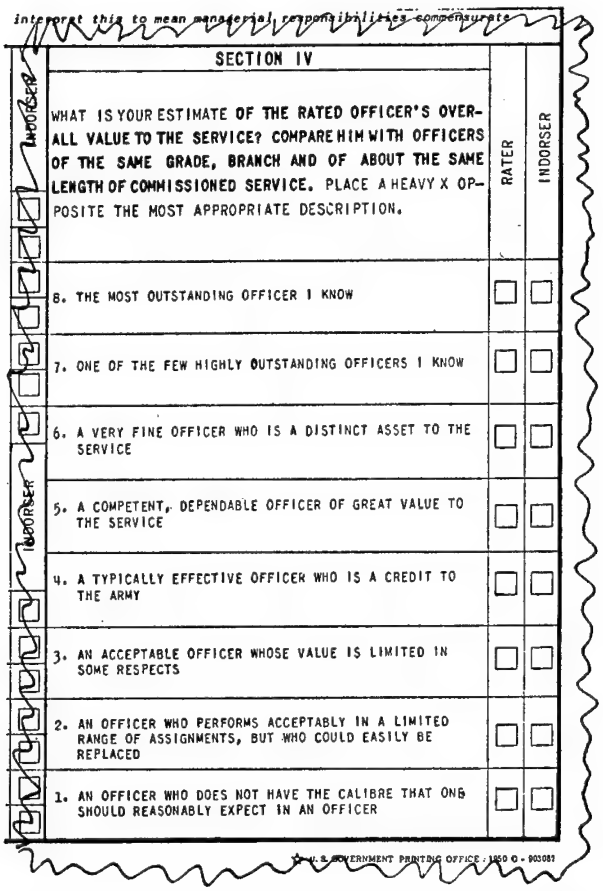
154. Rating Methods Suitable for Efficiency Reporting

The operating circumstances which permit use of such procedures as described above are so rare that the discussion here will be confined to procedures that are feasible in situations where superiors will be evaluating as few as one subordinate and which yield scores without the intermediate step of the judging process required by essays. Within these limitations, ratings can be considered as of two kinds—One is the traditional form on which the rater indicates how much of the rated element the ratee possesses. This includes check lists which may be considered measurement of "how much" on a two-point scale. The second type is known as the forced-choice method, in which the rater is asked not "how much" but which of two or more descriptions is most or least characteristic of the ratee.

a. Traditional Type of Rating Scales. The traditional type of rating scale has been used in almost numberless variations. One example occurs in section IV of the Officer Efficiency Report, DA Form 67–2, reproduced in figure 19. As noted above, the question asked the rater is one concerning the amount of an attribute—trait, characteristic, or over-all worth — possessed by the ratee. The traditional type of rating scale may concern general traits such as dependability or relatively specific traits such as ability to prepare a good report. The scale points may be purposely left somewhat vague, or they may be very carefully defined. The number of points on a scale varies from two upwards, but there usually are not more than seven or eight. All traditional methods are subject to the individual bias brought about by the human frailties discussed in chapter 3.

b. Forced-Choice as a Method of Efficiency Reporting.

(1) The question posed to the rater in this method is "Which of two or more descriptions applies most to the subordinate." This method can be considered an adaptation of a ranking procedure. Instead of placing in order of merit a



interpret this to mean managerial responsibilities commensurate

SECTION IV	RATER	INDORSER
WHAT IS YOUR ESTIMATE OF THE RATED OFFICER'S OVER-ALL VALUE TO THE SERVICE? COMPARE HIM WITH OFFICERS OF THE SAME GRADE, BRANCH AND OF ABOUT THE SAME LENGTH OF COMMISSIONED SERVICE. PLACE A HEAVY X OPPOSITE THE MOST APPROPRIATE DESCRIPTION.		
8. THE MOST OUTSTANDING OFFICER I KNOW	☐	☐
7. ONE OF THE FEW HIGHLY OUTSTANDING OFFICERS I KNOW	☐	☐
6. A VERY FINE OFFICER WHO IS A DISTINCT ASSET TO THE SERVICE	☐	☐
5. A COMPETENT, DEPENDABLE OFFICER OF GREAT VALUE TO THE SERVICE	☐	☐
4. A TYPICALLY EFFECTIVE OFFICER WHO IS A CREDIT TO THE ARMY	☐	☐
3. AN ACCEPTABLE OFFICER WHOSE VALUE IS LIMITED IN SOME RESPECTS	☐	☐
2. AN OFFICER WHO PERFORMS ACCEPTABLY IN A LIMITED RANGE OF ASSIGNMENTS, BUT WHO COULD EASILY BE REPLACED	☐	☐
1. AN OFFICER WHO DOES NOT HAVE THE CALIBRE THAT ONE SHOULD REASONABLY EXPECT IN AN OFFICER	☐	☐

U. S. GOVERNMENT PRINTING OFFICE : 1950 O - 903081

Figure 19. Rating scale on over-all value from Officer Efficiency Report, DA Form 67–2 (1 Sep 50).

group of individuals, it ranks traits within an individual. The method was applied to efficiency reporting in an effort to reduce the bias to which traditional ratings are subject. The reduction of such tendencies would result not only in increased validity but in increased fairness to the individual. Decisions affecting a career would be made in terms of differences in men rated rather than in terms of differences in rater standards.

(2) The logic of the forced-choice method is compelling. As a ranking method, it has the advantages noted in chapter 3 for this method. As yet, however, the effectiveness of forced-choice in efficiency reporting has not been fully established. This is in large part the result of the special problems involved in comparing the "validity" of ratings. This point will be discussed in considerable detail in paragraph 158.

It should be noted that other methods of applying the forced-choice principle are being sought because of the theoretical promise it has of reducing bias in ratings.

155. Efficiency Reporting is a Procedure

a. In efficiency reporting, there has been a strong tendency to emphasize the reporting form. If there is dissatisfaction with efficiency reporting, the initial reaction is likely to be—develop a new form. Actually, efficiency reporting is a procedure in which the form plays but one part, usually a minor one. It is, of course, no cure for inadequate observation and faulty judgment. At least four other factors are more important than the form of the report—

(1) Consistency of the procedure with its purpose.
(2) Policies governing the administrative uses to which the results are to be put.
(3) The administrative conditions surrounding the rendering of the report.
(4) Adequate observation and careful judgment.

b. The implications and importance of the first—consistency with purpose—have already been emphasized. The second and third have their major effect through influencing the attitude of the raters. The knowledge of the uses to which efficiency report scores are put and of the policies controlling these uses have a great deal to do with the attitude with which the rater approaches his task. If he has confidence that the scores are used with great care and due respect for their limitations, his desire to control the results of his rating can perhaps be lessened. The fourth point is so obvious as to need no elaboration.

c. The rater's attitude is also affected by the immediate conditions under which he renders the report. If those conditions do not permit sufficient observation of subordinates to allow for adequate rating, if the conditions prevent adequate time for careful completion of the report, or if those conditions require accomplishment of ratings for an excessive number of subordinates, it will be the exceptional case that careful, accurate ratings will be rendered. It

PAIRS

FOR RATER ONLY. For each pair of words or phases make a heavy X opposite the one that is the MORE DESCRIPTIVE of the rated officer.

1	A. Assigns men properly ☐ B. Keeps his word ☐	8	A. Maintains strict discipline ☐ B. Good educational background ☐
2	A. Courageous ☐ B. Respected by his subordinates ☐	9	A. Can select and define major objectives ☐ B. Thorough knowledge of his own branch ☐
3	A. Willing to take a chance ☐ B. Has physical endurance ☐	10	A. Temperate in his habits ☐ B. Self-confident ☐
4	A. Conscientious ☐ B. Takes action to correct faulty performance ☐	11	A. Is just ☐ B. Can get subordinates to attempt the impossible ☐
5	A. People seek his advice in personal matters ☐ B. Knows his subordinates ☐	12	A. Has a broad grasp of the problems ☐ B. Has foresight ☐
6	A. Alert ☐ B Has full knowledge of his job ☐	13	A. Vigorous ☐ B. Truthful ☐
7	A. Thoughtful planner ☐ B. Supports actions of subordinates ☐	14	A. Tenacious ☐ B. Makes practicable suggestions ☐

OR

TETRADS

This section consists of sets of phrases which describe characteristics related to proficiency on the job. Consider each set and decide which phrase is MOST DESCRIPTIVE of the officer being rated, and which phrase is LEAST DESCRIPTIVE of the officer. Judge the sets independently -- it is not necessary to be consistent, as you are describing the officer, not evaluating him. Blacken in the space to the right of the MOST DESCRIPTIVE and LEAST DESCRIPTIVE phrases in the appropriate column.

	MOST	LEAST
1	A. Becomes dogmatic about his authority.	
	B. Careless & slipshod in attention to duty.	
	C No one ever doubts his ability.	
	D. Well-grounded in all phases of Army life.	
2	A. Follows closely directions of higher echelons.	
	B. Inclined to "gold-brick."	
	C. Criticizes unnecessarily.	
	D. Willing to accept responsibility.	
3	A. A go-getter who always does a good job.	
	B. Cool under all circumstances.	
	C. Doesn't listen to suggestions.	
	D. Drives instead of leads.	
4	A Always criticizes, never praises.	
	B. Carries out orders by "passing the buck."	
	C. Knows his job and performs it well.	
	D. Plays no favorites.	
5	A. Constantly striving for new knowledge and ideas.	
	B. Businesslike.	
	C. Apparently not physically fit.	
	D. Fails to use good judgment.	
6	A. Cannot assume responsibility.	
	B. Knows how and when to delegate authority.	
	C. Offers suggestions.	
	D. Too easily changes his ideas	

Figure 20. Forced-choice phrases may be assembled in pairs or tetrads.

is extremely important, then, that when management is convinced of the need for an efficiency reporting procedure, this conviction be expressed in the administrative conditions surrounding the rendering of the report.

d. To repeat, the rating form can be an aid in making more accurate and useful judgments by requiring the judge to be more careful, more critical in the recording of his judgment or by providing information which might be used to evaluate the worth of the judgment. No matter what the form is, the far more important aspects of rating are the procedures employed and the competence of the judge as a rater.

Section IV. MAJOR PROBLEMS IN EFFICIENCY REPORTING

156. Acceptability of the Rating Procedure

a. What has been said just above points to the importance of the rater's attitude toward the rating procedure. At best, rating is a difficult and disliked task. The more the raters accept the system, the greater the likelihood that better reports will be rendered. This problem of acceptability has usually been discussed in terms of "selling the scale" or the "need for selling the

raters." It has already been made clear that more is involved than salesmanship. Those concerned need to be convinced of at least three points—the need for a rating system, the ability of the rating system to meet this need, and the justice of the policies which govern the use of the rating scale. Further, raters cannot be expected to believe efficiency reporting is important if higher echelons do not make some effort to provide the conditions and the time that makes good reporting possible.

b. There is even more involved. A supervisor generally feels, and with some justification, that he is in the best position to control action with respect to his subordinates. He therefore wants to know where he is placing a subordinate on a scale. What the headquarters responsible for the personnel action needs to know in evaluating a particular man is—How many others are there like him? As a result, each rater's scores must be converted to show the ratee's standing in his competitive group. As will be seen when the problem of standardization is discussed, it is impossible to tell the rater the competitive standing of the ratee. As a result, the rater is likely to feel that information is withheld. Solution to a problem of this kind is not easy. Long-term educational programs appear to be the only method that has any promise at all.

157. Rater Tendencies and Efficiency Reporting

Three persistent and outstanding findings have emerged from Army * research studies on ratings: (1) tendency of a rater's evaluations of different aspects of an individual's performance to be highly related, even when the various aspects appear to be logically independent; (2) the unsatisfactorily low agreement among raters; and (3) the strong tendency toward leniency in efficiency reporting.

a. High Relationships Among a Rater's Evaluations of Different Aspects of Job Performance.

(1) As mentioned above, raters tend to rate an individual much the same on different traits or aspects of job performance. In other words, their ratings on supposedly different aspects of performance turn out to be highly correlated. The tendency ("halo") has already been discussed in chapter 3 with reference to criterion ratings. It operates, probably to a greater degree, where efficiency reports are concerned. The rater's desire to control the administrative actions he believes will result from his report makes this almost inevitable.

(2) From the point of view that what is desired in efficiency reports is a measure of over-all worth, the tendency toward high correlation of ratings given by the same rater for one person may not be of great consequence. This is true, however, only if those who use the ratings do not attempt to make differential assignments on the basis of them. The tendency is very unfortunate from the standpoint of research. If ratings on more specific characteristics could be accurately obtained, they could be combined to yield perhaps a better measure of over-all worth.

b. Low Agreement Among Raters. Different raters are likely to observe the ratees in different situations and hence base the ratings on different samples of behavior. Raters may interpret the scale points differently; they may have different professional, personal, and social relations with the ratee; or they may observe the ratee in the same situations but have different standards or even different values for the significance of the behavior for over-all worth. Much of the lack of agreement can be attributed to rater biases of this sort. Error of an unsystematic nature, of course, still accounts for a certain amount of the disagreement. The implication of this disagreement for efficiency reporting is that it is advisable to use an average of as many ratings as possible prior to taking personnel action. It is usually impractical to get a number simultaneously, but successive ratings can be averaged.

* The word "Army" is emphasized in this context, particularly with reference to rater agreement. Some studies in industrial settings report higher rater agreement than is typical for Army studies. Longer opportunity to observe on the same job—up to years, in some instances—may be an important factor in this greater agreement. Further work is needed to reconcile the different findings.

c. Leniency of Raters.

(1) It is well established that raters rendering efficiency reports tend to be more lenient than raters accomplishing nonadministrative ratings. When giving a rating that is used for personnel decisions, it is natural to give a subordinate the benefit of the doubt. This tendency makes the logical or absolute scale (for example, the adjectival scale of the former Efficiency Report, Form 67) different from the scale based on the relative positions in a competitive group. This has serious consequences for standardization problems (par. 159).

(2) The problem of leniency is complicated by the fact that all raters are not lenient to the same degree. These differences may be larger in efficiency reports than for criterion ratings and may be one of the reasons it is difficult to increase the validity of efficiency reports.

d. Concentration of Ratings.

(1) Another way of describing the tendency to be lenient is to say that raters tend to give a disproportionate number of high ratings. This lack of dispersion of ratings can be a serious problem limiting the administrative use of ratings. Army-wide efficiency reports are of little value to a remote headquarters responsible for personnel actions if the ratings of large numbers of men are so close together that there is no way of distinguishing relative degrees of competence.

(2) One of the reasons for the lack of adequate dispersion may be the confusion over reference points. Raters may be viewing the low end of the scale as the low end of the general population instead of the low end of the competitive group. The rating should be directed at the competitive group and not the general population. Concretely, an officer compared with other officers of similar grade may be the poorest of the lot and still be considerably better than the average of the entire male population of the country. In rendering administrative ratings, the reference point should be the average of the competitive group, not the average of the general population.

158. Validating Efficiency Reports

a. Ratings are so frequently used as criterion measures that the question, "How can ratings be validated?" appears almost paradoxical. The problem of validating efficiency reporting procedures received only passing attention until toward the close of World War II when the Army began intensive research on efficiency reporting methods. Most previous studies were concerned with problems of distribution, reduction of halo, rater standards, and reliability. The possibility of using a number of carefully collected anonymous ratings from associates (superiors, equals, and subordinates) as a criterion for efficiency reports was recognized. The single efficiency report that produced scores which agreed best with the criterion average could be considered the least biased or the most valid.

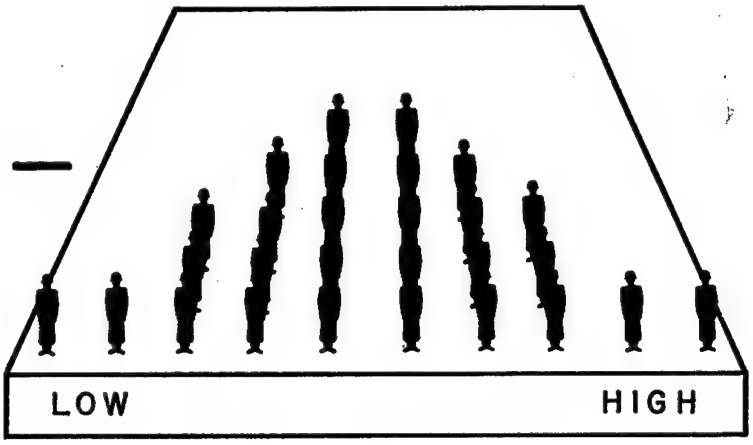
b. While this approach to the validation of efficiency reports is sound, there are several problems that need to be solved. One concerns the kind of rating that should be used in the criterion. In the use of associates' ratings as criterion measures for efficiency reports, it has been observed that the type of rating that was most like the type used in the criterion tended to have the highest validity coefficients. This suggestion that like scales correlate with like scales is referred to as "technique contamination." It can be particularly serious when two techniques as different as the forced-choice and traditional rating scales are being compared.

c. It has long been recognized that the rating being validated and the rating used as criterion should not be made by the same raters. If the raters were the same, the two sets of ratings would show agreement just because the same raters made the two sets of ratings ("rater contamination"). Thus, in validating a rating procedure, it appears necessary to design the study so as to reduce both technique contamination and rater contamination.

RATINGS

SHOULD BE DISTRIBUTED—

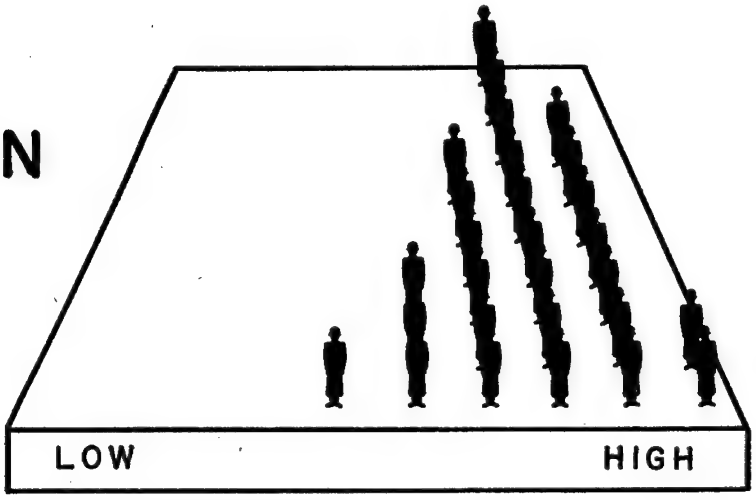
LIKE THIS —



BUT OFTEN

LOOK LIKE

THIS —



BECAUSE RATERS TEND TO

1. CONCENTRATE THEIR RATINGS
2. BE LENIENT

Figure 21. Two important difficulties in obtaining valid ratings.

d. In view of the discussion of rater tendencies and criterion problems, it appears necessary to measure the validity of efficiency reports under conditions that approximate the official operating conditions.

e. In conclusion, it can be said that the problems involved in validating ratings of usefulness to be used in personnel actions (efficiency reports) are different in many respects from those involved in validating other instruments. The need for validation, however, remains the same.

159. Standardizing Efficiency Reports

a. Just as was the case in validating efficiency reports, standardizing them presents problems not involved in standardizing tests. In the case of tests, the primary purpose of converting to a percentile or standard score scale is to permit interpretation of the score in terms of a comparison with a standard reference population and to permit comparison of standing on several tests taken by the same individual. In the case of officer efficiency reports, a standard scale is needed to compare different reports obtained during an officer's career. Even more important is the desirability of averaging the reports, which would not be possible unless the same standard scale were used for the different reports.

b. A serious problem in terms of making an efficiency report acceptable to the rater is involved in the standardization of an efficiency report. When a test raw score is converted to 120 on a scale of 51 to 150, with the meaning that it is exceeded by 16 percent of the population, the fact is accepted with little comment. When an efficiency report is converted into such a scale, the rater often feels that the score does not represent his intentions. This is far more frequently the case when the standard score is below average, that is, below 100. This difference between relative and absolute meaning of efficiency report scores is one of the basic problems in the acceptability of a report. The problem imposes a dilemma. Raters, at least in the Army, desire to know where on a scale they are placing a subordinate, whether it is below or above average. If, however, this information is divulged, the rater is influenced by it and the standardization no longer holds. Further, if the rater rates in terms of estimation of such relative position, he will not be rating in terms of whatever content is included in the rating scales. All systems would reduce to the question of percent position in a group.

c. Raters can, therefore, be given only the information that was available when the original standardization took place, and they should be told why this is necessary. Otherwise, the standard scale will no longer be appropriate. Further, it is apparent that standardization must be in terms of the operating situation. It cannot be done on an experimental basis.

d. As a matter of fact, one of the reasons for frequent changes of efficiency report forms is this tendency for raters to desire to control the ultimate placement of their subordinates. All Army efficiency report scores have shown a strong tendency to increase on the average from year to year. As officers learn about the relative positions their subordinates attain as a result of their rating, they tend to become more lenient. Rating scores rise, and the standardization becomes obsolete, as does the report form itself.

160. Improving Efficiency Reporting

a. Averaging Reports.

(1) It is well-known that averaging a number of ratings for a single ratee will reduce the effect of bias shown by the individual rater. Even if certain techniques are found to be effective in reducing the bias in a single report, averaging several such reports will yield a still better measure. This principle has been consistently applied in obtaining ratings for criterion purposes, where the score used is the average of ratings by a number of raters. The principle has been recognized also in the systems of officer efficiency evaluation established by the Army. The General Efficiency Rating (GER) used earlier and the present Over-all Efficiency Index (OEI) both involve methods of combining an officer's successive efficiency report scores into a long-term (5-year) composite score.

(2) Even though averaging reports makes a marked improvement in a system, there remain compelling reasons for improving single efficiency reports. Decisions frequently have to be made concerning retention of an Army officer before he has had the opportunity to accumulate a series of reports. Further, Army assignment is such that the number of reports an officer has received varies considerably. Some officers, because of school assignments, may receive but one report over a 2- or 3-year period. This report then should be as accurate as it can be made.

(3) In the discussion of criterion ratings in chapter 3, several ways of improving single ratings were discussed—educating the rater, improving the conditions under which ratings are obtained, making a better selection of raters, use of the forced-choice technique, and greater care in obtaining the rating ("guided" rating). These same factors apply in varying degrees to efficiency reporting. One needs to be added—Clarification and dissemination of policy controlling the use of efficiency reports.

b. Clarifying Purposes of Efficiency Reports. This is a management problem. Confusion of purpose, it has been pointed out, can result in ineffective reporting. A clear conception of the purpose of the reporting, and keeping conditions and requirements consistent with this purpose, can go a long way toward improving an efficiency reporting system.

c. Policy Controlling Use of Efficiency Reports. This is another problem for management. It concerns the improvement of efficiency reports because it can have a marked influence on the willingness to render objective reports.

d. Education of the Rater. Perhaps the most important additions to what was said in chapter 3 on this topic are (1) the need for conveying to the rater information concerning the purposes of the rating and the policies governing its use; (2) the necessity of making the rater aware of the important factors to observe prior to rating; and (3) the value of actual supervision and guidance of the rater, just as in any other training program. The importance of supplying the rater with information concerning policy is apparent from the influence of such knowledge on the attitude with which the rater approaches his task. Need for training in rating is likewise apparent. It need be added only that the Army faces a special problem in training and educating raters. Its size precludes direct contact with all raters. A practical way of indoctrinating raters would appear to be the inclusion of appropriate material in curricula of various key schools. The repeated emphasis that this permits would appear to offer promise for a real improvement in the Army efficiency reporting system. However, this form of education would need to be supplemented by the supervision of raters in making official ratings. This supervision could be furnished by the rater's supervisor as part of his responsibility in improving his subordinates.

e. Administrative Conditions and Efficiency Reporting.

(1) Administrative conditions can influence the quality of efficiency reporting both by creating conditions conducive to careful rating and by affecting the attitude of the rater toward his task. Some examples of conditions having the possibility of such influences are —the time allotted for the rating process, the number of subordinates for whom a particular rater must render a report, the frequency with which reports must be rendered, designation of conditions when reports need not be rendered, and selection of raters.

(2) Attention to such factors as these would undoubtedly improve efficiency reporting procedures. But one of them —selection of raters—needs further comment here. Selecting raters according to characteristics which make for good rating is a relatively new field for research. Once characteristics that make for good rating can be identified, efficiency reporting could be improved by enabling commanders to designate the more effective raters, at least where options exist.

Section V. SUMMARY

161. Ratings Used for Administrative Purposes

a. Ratings are used to estimate over-all value.

b. Rating procedures may be intended to accomplish different purposes. Efficiency reports are ratings for administrative purposes.

c. Efficiency reports are needed in the Army to provide estimates of over-all value of men in a large competitive group.

d. Kinds of efficiency reports—

(1) Unsuitable—essays, ranking, paired-comparison, guided ratings.

(2) Suitable — traditional rating scales, forced-choice.

e. Efficiency reporting is a procedure, not a form.

f. Major problems in efficiency reporting—

(1) Acceptability of method.

(2) Rater tendencies which reduce usefulness of report—

(*a*) Similarity of ratings on different aspects.

(*b*) Low agreement among raters.

(*c*) Leniency of raters.

(3) Criterion problems in validating efficiency reports—

(*a*) Use of associates' ratings.

(*b*) Technique contamination and rater contamination.

(*c*) Need for validating under conditions approximating normal operations.

(4) Standardization of scores—

(*a*) Need for converting report scores to show standing in competitive group.

(*b*) The conflict between the rater's intended score and the relative score in the competitive group.

g. Improving efficiency reporting—

(1) Averaging a number of reports.

(2) Clarification and dissemination of policy.

(3) Education of raters.

(4) Improving conditions under which reports are rendered.

(5) Selecting raters where options exist.

(6) Rating techniques.

CHAPTER 11
THE ADMINISTRATION OF ARMY TESTS

Section I. INTRODUCTION

162. General

The scientist develops tools; the technician puts them to use. The instruments devised by personnel psychologists are constructed on the basis of extensive knowledge, and carefully standardized. As such they are valuable means of increasing the accuracy of observations or personnel measurements, and of revealing important and useful information. Once an instrument is constructed, great care must be taken that scores are obtained with these instruments exactly as specified by the research which built the instrument. The aim of the present chapter and the one following is to clarify the work of the technician in administering and scoring personnel evaluation instruments so that results of the greatest possible value to the Army will be produced.

163. Importance of Test Instructions

Specific directions for administering and scoring are set forth in the manuals which accompany each Army evaluation instrument and are as much an integral part of every test as the questions themselves. Included in specific instructions are the exact wording of the directions to the examinees, the time limits and directions for scoring, and descriptions of the proper technique for recording and interpreting results. Deviations from any one of these can affect the accuracy of measurement. Manuals of directions must be adhered to strictly.

164. Testing Situation Must Be Standard

Since it is the function of every evaluation instrument to compare each individual with others in the Army population, it follows that the condition under which tests are administered and scored must be the same for every soldier, regardless of when or where the test is given. The scores of men who are tested in noisy surroundings or by slipshod methods are not comparable to those of men examined under favorable circumstances. Nor are the scores in all probability accurate indications of the real abilities of those men. The use of such scores can only result in incorrect evaluation and use of results with attendant loss of efficiency to the Army.

a. Testing conditions and procedures should be so standardized that if it were possible to find two individuals exactly alike, both would achieve the same scores, though tested at different times and in different places. Only scores obtained under standardized conditions can be relied upon to show what may be expected of men.

b. No valid comparisons can be made between the scores of men performing at different levels of motivation. Thus every effort should be made to insure that all men perform to the best of their ability. Standard conditions should therefore be optimal conditions.

c. Tests should be administered and scored in a manner identical with that employed in their standardization. Standardization (ch. 2) involves the administration of each test to a sample group of men with known characteristics in order to obtain norms by means of which each subsequent score may be evaluated and interpreted. In other words, test performances in the field are evaluated by comparing them with the performance of the men in the

standard reference population. If this comparison is to be a valid one, the administration and scoring should be identical in the two instances. Army tests are always standardized in the field under conditions which can be duplicated in field installations. The principles set forth in this chapter make it possible to duplicate the standardization conditions.

Section II. PRINCIPLES AND PROCEDURES FOR ADMINISTERING GROUP TESTS

165. General

It has already been stated that the procedures for administering tests should be such as to call forth the best performance of which the individual is capable under standard conditions. Each individual will tend to do his best if his environment is reasonably free from distracting influence, if he understands what he is to do, and if he considers it worthwhile to do his best. The first of these conditions depends upon the physical aspects of the testing situation; the others upon the techniques employed by the examiner in controlling the testing situation.

166. Physical Surroundings

All behavior, including test performances, takes place in an environment. Since it is impossible to administer tests in a "vacuum," the next best thing is to take steps to insure that the environment provided is standard for all administrations of the test, and that it does not impede or hamper the performance of the examinee. While it is recognized that ideal testing conditions cannot always be achieved with the limited facilities available in field installations, attention to the following factors should provide conditions that are adequate in most cases.

a. So far as possible, the testing room should be quiet. Noise is one of the principal sources of distraction from concentration and mental effort. Yet, absolute silence is neither necessary nor desired. The individual who can perform satisfactorily only in a soundproof room is going to find few places in the Army suited to his peculiarities. Noise which continues steadily at a moderate and fairly even level of intensity can be considered as normal for testing conditions. Such noise would include the steady hum of indistinguishable voices from another part of the building, the drone of machines, or the continuous but muted clatter of typewriters. But a sudden shout outside a window, a bell, the clatter of unloading a truck, the blare of a radio, or the sound of persons passing through the room are very distracting to examinees.

b. Consideration should be given to the acoustics of the testing room. The examiner's voice must be clearly audible to all men being tested. The public address systems now found in most Army testing rooms have solved this problem, but not without adding others. Care should be exercised in placing loudspeakers and in locating microphones. The public address system is still something of a novelty, and people feel an urge to see where the voice is coming from. An invisible ghost-voice can cause considerable craning of necks and unnecessary distraction; so, if a test must be given from an unseen location, a preliminary announcement to this effect will dispel distracting curiosity. The level of amplification should also be controlled. Loud directions booming forth above one's head can be very disconcerting.

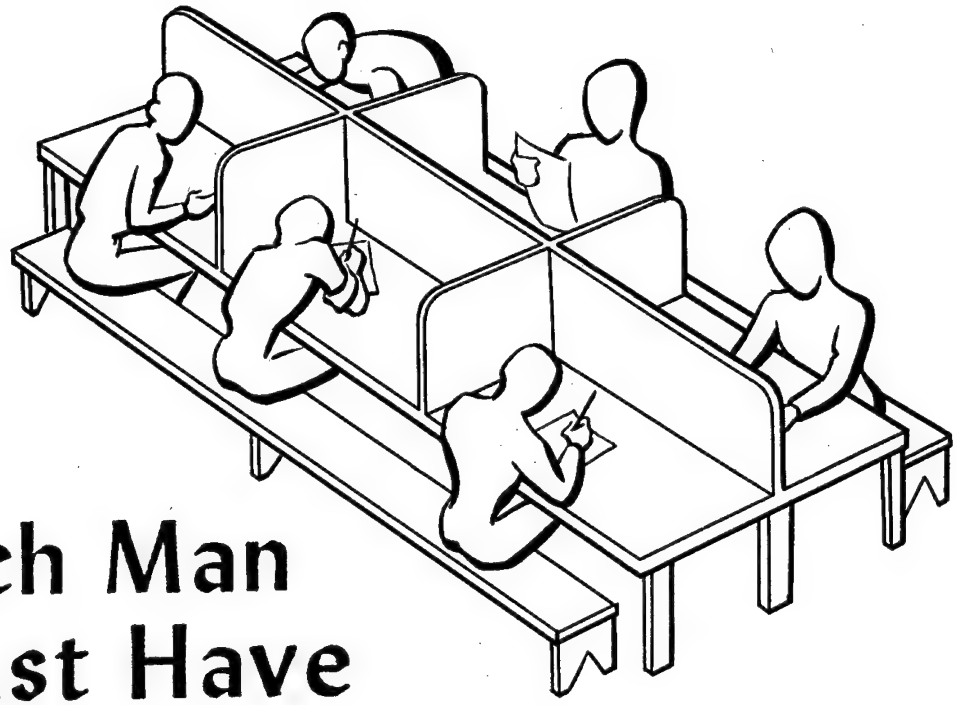
c. The testing room should be well lighted and ventilated. There must be sufficient illumination on the working surface to prevent eye strain. If a light meter can be obtained, the illumination in various parts of the room should be checked. Illumination of the working space is the important thing. A light meter laid on this space should register 6 to 10 foot-candles. Special care should be exercised to avoid glare spots and shadows; there is perhaps nothing as annoying as having part of the test paper intensely illuminated with the rest in the shadow cast by a pillar, a partition, or the examinee himself. Conditions of temperature, humidity, and ventilation are sometimes difficult to control, yet every effort must be made to do so. No one can perform at his maximal efficiency in a room where the air is hot, sticky, or stale.

d. Among the foremost factors of importance are the spatial arrangements of the testing room. If conditions permit, the examiner should be provided with a raised platform or

TO GIVE HIS

Best Test

PERFORMANCE



Each Man Must Have

1. A good test environment
2. A complete understanding of directions
3. A desire to do his best

Figure 22. Factors affecting test performance.

rostrum in a part of the room where he can see, and be seen, by all men being tested. This is especially important where test directions call for the presentation of charts or other demonstration material. The desks or tables for the examinees should be arranged to leave aisles for the proctors to use in distributing and collecting tests materials and in circulating about the room during the test. If possible, there should also be enough space between rows to allow passage. Otherwise, when it becomes necessary for a proctor to reach an individual in the middle of a row, there is much treading on toes and knocking elbows en route. Such distractions need no further comment. The writing surface itself should be flat and smooth and free from cracks. Pencils have an irritating way of punching through the answer sheet when there are knotholes and cracks in the board beneath. If available tables are rough, a tight covering of linoleum or press board (masonite) will correct this. The space allotted

to each individual must be wide enough to accommodate both a test booklet and a separate answer sheet. Chairs with writing arms should not be used for testing since the writing surface provided is far too narrow. Many installations now are equipped with large tables with vertical partitions separating the surface into booths approximately 30 inches wide and 18 inches deep. The partitions insure each person sufficient room and prevent the overcrowding of timid souls by neighbors with aggressive elbows. They also discourage community collaboration. If tables like these are not available and cannot be constructed, mess tables make an adequate substitute. However, if the mess tables are still being used for eating, the hours just before and just after meals should be avoided in the testing schedule. The noises and odors issuing from the kitchen or the clatter of dishes from another part of the mess hall fall into the category of distraction.

e. The temptation to give or to receive aid always seems to be present wherever people are examined in groups. Aside from the fact that cheating is reprehensible from the viewpoint of military discipline, its effect on the validity of the test score requires that it be prevented. For classification purposes, the Army is interested in how many correct answers the individual can obtain by himself, not how many he can copy from his neighbor. The use of partitioned booths (described above) or of alternate seating will help to prevent collaboration. In addition, all blackboards and charts in the room should be checked to insure that no material is left visible to help the examinee, and all test booklets which are to be reused should be examined after each session and any answers written therein erased. Despite all precautions, the proctors will still have to prevent cheating during the examination. For this reason, proctors should circulate (as quietly as possible) rather than remain at a fixed post. The mere nearness of the proctor on his rounds is often a sufficient curb on cheating.

f. Not all distracting influences are in the external surroundings. The condition of the individual, his physical and mental state, also affect his test performance. The man, for example, who has just had disturbing news from home, or is in physical distress, is in no condition to do his best on an examination. In individual cases, these factors cannot always be foreseen, but for the group as a whole, much can be done by scheduling testing sessions at a time of day when fatigue or physical or emotional discomfort can be expected to be at a minimum. In normal circumstances, the morning hours will be the best time to schedule an examination and the end of a long day the poorest. Where possible, activities should be controlled so as not to interfere with testing schedules. In the reception center, for example, processing should be so regulated that testing does not follow hard exercise, long hours of waiting in line, or immunization "shots." In all cases the test officer, examiner, and proctors should be alert to the signs of genuine distress, and the affected persons should be excused until a more propitious occasion.

167. Testing Session

The ideal testing session is a smooth-running organized affair. Since its primary purpose is to obtain reactions to standard questions under standard conditions, the major portion of the time is allotted to taking the test itself. All other activities, such as assembling and seating the men, distributing materials, giving preliminary directions, and collecting materials, are necessary adjuncts to the main event. Yet it is the management of these details which can make the session a smooth-running operation or chaotic confusion—an ordeal to both examiner and examinee. The secret of success is control. If the examiner is at all times master of the situation, he can keep things moving and organized. But if the examiner is uncertain and stumbling, if there are unnecessary delays, the group will become restless and irritable. Control is best achieved by careful preparation and practice in all phases of the process—preliminary arrangements, test administration, and the collection and disposition of materials. The discussion that follows will cover general principles and specific suggestions for making the testing session an orderly, systematic affair.

168. Preparation for the Testing Session

a. General. Preliminary planning for the testing session involves the careful selection of

GOOD TEST ADMINISTRATION REQUIRES A GOOD TEST ENVIRONMENT

Figure 23. Some conditions detracting from good test environment.

the testing team, the instruction of all members, and practice drill in all required testing procedures. The examiner is selected for the quality of his speaking voice and for his ability to handle groups of men. While no one demands that the examiner have the speaking voice of a trained actor, he should have one that can be understood easily. His accent should either be indigenous to the group being tested; or, if the group is from many parts of the country, his accent should be "standard American"—the accent common to the Midwest. It is also desirable that the examiner be capable of controlling the testing situation—in many cases, noncommissioned grade gives the examiner the prestige necessary. Good testing administration demands—

(1) Thorough preparation.
(2) Following directions exactly.
(3) Accurate timing.
(4) Care of test materials.

b. The Examiner's Preparation. The examiner should make a careful study of the manual to make sure that he knows the purpose of the test, the materials needed to give it, the directions to be read, and the problems which are likely to arise. He should study those directions which are to be read aloud until he can read them in a normal manner, without stumbling over unfamiliar words, losing his place, dawdling, or racing through in an unintelligible patter. Familiarity with the contents of the test itself is also invaluable. It is excellent practice for both the examiner and all proctors to take each test in the normal fashion before attempting to administer it; this procedure should be standard whenever a new test is installed or new examining personnel are trained. In this way, the examiner gains an appreciation of the men's viewpoint on the test and learns how to anticipate, and thus be prepared for, the common questions which may arise.

c. Duties of the Proctors. The examiner is responsible for instructing the proctors thoroughly in their specific duties. The common practice of snatching any man not at the moment occupied, and making him a proctor, then and there, should be frowned upon. It is far more efficient to designate regular testing teams responsible for the administering, proctoring, and scoring of tests. Each proctor should be assigned a certain section of the room for which he will be responsible. Before the testing period, he should check the materials to be used to make sure that they are all there in good condition and order, and in sufficient quantities. He should know the order in which these materials are to be distributed and collected, so that, when the time comes, he can execute this phase of his assignment with efficiency and dispatch. With the administration of the test itself, his real job begins. While directions are being read and while the test is being taken, he should patrol his assigned area. Within this area, he is responsible for—

(1) Seeing that each examinee has all the necessary materials for taking the test, and furnishing these, especially pencils, where needed.
(2) Insuring that each examinee is following the directions correctly and understands what he is to do and how he is to do it. The proctor should be alert to detect incorrect methods of marking answers where separate answer sheets are employed.
(3) Seeing that each examinee is doing his own work, independent of his neighbors.
(4) Excusing from the examination any person who is or becomes too ill to continue without discomfort.

169. Administering the Test

All Army tests should be administered strictly in accordance with the manuals of directions which are supplied with them. A test given without the aid of the manual is no more trustworthy and reliable than a rifle without sights. What follows here can be considered as instructions on how to use the manual in administering and timing a test.

a. It has been previously said that the value of any test score will depend on the extent to which the examinee understands just what he is to do and the degree to which he considers it worthwhile to do his best. The examiner's primary responsibility, in fact his main function, is to elicit this willingness to work and

to provide the proper instruction. The first is a matter of the appropriate stage setting and of favorable attitudes. The second is handled by oral directions contained in the manual.

b. These oral directions include, among other things, a brief informal statement explaining the test to be given, how the results will be used, and why it is important for each person to do his best on it. The aim of these remarks is a difficult one; to dispel anxiety and release tension, yet at the same time, to stress the necessity for maximal effort and output. A careless presentation may create the impression in the minds of the men that the test they are about to take is of no consequence and in no way related to their future Army careers. Or it may so impress them with the seriousness of the situation as to give rise to disturbing tensions. On the whole the best results will be achieved through a brief, straightforward but nontechnical statement of facts, delivered in a manner that is neither formidable nor severe, nor yet so jocular or perfunctory that it defeats its own purpose.

c. Having set the stage and gained the necessary cooperation, and having distributed the test material, the examiner next informs the men exactly what they are to do. There is only one way to do this—by reading aloud the directions provided in the manual. And it means reading aloud all the directions that are to be read aloud and no more. The test administrator should not make the mistake of reading aloud to the examinees any of the explanatory statements in the manual which are intended solely for the test administrator. Directions should be read in a natural voice, in a smooth coherent fashion. Hence, they must have been thoroughly practiced. Notice, however, that they should be *read,* not paraphrased, given from notes or memory, or adapted to someone's idea of what is more appropriate for local conditions. Every test was "zeroed in" with its directions; the sight setting must not be altered or the results will be wide of the mark.

d. Many Army tests are given with certain time limits which must be strictly observed if testing conditions are to be uniform from session to session, and from place to place (par. 81). These time limits, either for the complete test, or for various parts of the test separately, are always specified in the test manual. They are exact, not approximate; so timing should be handled with care. If a stop watch is available, it should be used. If not, any good watch with a second hand will serve, if used in the following manner:

(1) When giving the signal to start the test, note down on paper the hour, minute, and the second of starting.

(2) Write below this time the hours, minutes, and seconds of working time for the test as specified in the manual.

(3) Add these two figures to obtain the exact time when the signal to stop work should be given. (If the minutes add up to more than 60, of course, the 60 minutes would be carried as an additional hour and the excess listed as minutes.)

Example:

Starting time	1451:00
Time limit for test	45:00
Stopping time	1496:00
	or
	1536:00

The signal to stop should, therefore, be given promptly at 1536. The timing should always be done in this way. It is unwise to trust to memory or to attempt the necessary computations mentally. And it is good practice to have some of the proctors check the timing independently.

170. Collection and Disposition of Test Materials

After the signal to stop work has been given, the materials should be collected as quickly as possible. The period between tests or at the end of a session can be one of tremendous confusion with everyone talking, comparing notes, reaching for coats and hats, and being anxious to leave. Under these circumstances the test materials are likely to be collected in haphazard fashion, with booklets and answer sheets jumbled together. In the confusion booklets have a way of disappearing. Army tests are classified RESTRICTED and must be accounted for. It should be remembered also that the booklets will be used again and the answer sheets

have to be scored. Planned and orderly collection of materials will pay big dividends in the time saved during future operations. The following system achieves a maximum of order and control.

a. As soon as the stop signal is given, the examiner should instruct the men to remain quietly in their seats and to follow directions in order to expedite the collection of materials.

b. He should then direct them to pass these materials to the ends of the rows, specifying which end. The materials should be passed separately, first answer sheets, then test booklets, then supplementary materials such as scratch paper, and finally, pencils.

c. The men at the ends of the rows should be instructed to stack the materials in separate piles, making sure that all booklets are closed, with the cover sheet outside, and that all answer sheets are faced the same way.

d. When this is done, the proctors can collect the materials and at the same time make a rapid count of the numbers turned in. Only after all materials have been checked in and accounted for should the group be dismissed or the next test begun.

e. An alternative method is to instruct the men to bring all of the testing materials to the front of the room and place each item in an appropriate pile on a table before leaving. In this way proctors can insure that each examinee turns in all of his testing materials before he is permitted to leave.

f. Frequently some of the examinees will give up or will finish before the time limit is up. Care should be exercised to insure that these men do not leave until they have complied with all directions in the manual for administration of the test and that they have turned in all of their testing materials.

171. Care of Booklets

After each testing session, and certainly before the next session, all test booklets should be carefully scrutinized for answers or marks of any kind. In spite of all warnings, some persons will write answers in the booklets or use them for scratch paper. If the answers or marks can be erased, this should be done. If not, or if the booklet is worn or torn, it should be destroyed in accordance with proper procedures for handling classified material. All used scratch paper should be destroyed. All tests and testing supplies should be kept secure according to the appropriate security classification when not in use.

172. Answering Examinees' Questions

As a rule, questions asked by examinees prior to the actual time of beginning the examination are answerable. These questions usually pertain to testing procedures, time limits, and the purposes and uses of tests. After the examinees have begun the test, test administrators and proctors should be especially careful to avoid giving away information that might influence the proper evaluation of an individual, as in the following circumstances:

a. If an examinee asks the test administrator or proctor to tell him the answer to a specific test question, the answer to him should indicate that it is not permitted to tell the examinee the correct response to a test question since that would not allow the test to do an accurate job of evaluating him.

b. If an examinee states that he does not understand what a question means and asks the test administrator or proctor to explain it to him, he should be told that to do so would influence his responses and thus not permit the test to give a true evaluation. He should also be encouraged to do the best he can.

c. Even though some of the questions presented by the examinee seem "stupid" or appear to be repetitious, impatience or contempt should not be displayed by the test administrator or proctor. In fact, an unusual number of questions may be an indication that the test administrator is not following directions, is not speaking loudly or clearly, or that acoustics are unsatisfactory. If such is the case, corrective action should be taken.

d. In general, the examiner should try to answer as many questions as possible without giving away answers or providing clues to answers.

Section III. ADMINISTERING INDIVIDUAL TESTS

173. General

For the most part, evaluation instruments employed by the Army are of the paper-and-pencil type administered to groups of examinees at one time. In certain special circumstances, proper evaluation of the soldier will necessitate the administration of an individual test. The expenditure of time and effort is greatly increased with the employment of such instruments, and they should be used only when specified by appropriate regulations. In general, the individual type of test is recommended for cases in the following categories:

a. Where the paper-and-pencil type test is inappropriate because the subject is lacking in the educational skills of reading and writing.

b. Where more personal contact is needed to insure that the examinee is at ease, is properly motivated and encouraged, and knows just what he is supposed to do.

c. Where it is desired not only to determine the individual's over-all score, but also to give the examiner the opportunity to observe him at work and to estimate his specific strengths and weaknesses.

174. Individual Testing Session

The individual testing session is a more personal and somewhat less formal affair than the group testing session. This does not make it easier to manage; on the contrary, the individual test administrator needs much more than average training and experience to achieve results which can be accepted with any confidence. His task is a difficult one. He must administer a rigidly controlled and carefully standardized set of problems under conditions precisely as specified, all the while creating an impression of friendly informality. Men assigned to this job must be selected with care. The ideal examiner is a man with a knowledge of the principles of psychological measurement and an appreciation of the needs for exactness and precision. He is personable and friendly and an easy conversationalist. He is patient and tolerant, never given to a show of arrogance, flippancy, or sarcasm, no matter how absurd the responses of the subject might be.

a. Individual tests, because they are essentially personal interviews, should be given in an atmosphere of privacy. Special rooms are recommended, but separate booths divided by partitions will serve where facilities are limited. Examiner and subject should be provided with comfortable chairs facing each other across a table or desk. This table or other working surface should be large enough to accommodate all of the test materials and provide room for the examiner to jot down responses on the record sheet. The field table is of the proper dimensions for most individual testing.

b. Before undertaking the administration of an individual test, the examiner should make a careful study of the manual of directions and of the test materials. The exact wording to be used in presenting the materials will be specified, and should be rehearsed until it can be read in a normal conversational manner. The examiner should also practice the things he is to do—the placement of the materials, movements, pointings, demonstrations, etc.—until these are smoothly coordinated with the verbal directions. Finally, he should give practice administrations with "guinea pig" subjects, under the supervision of a qualified examiner until he can maintain the subject's interest and confidence, select and use necessary materials and instructions without fumbling, and make of the whole procedure a smooth and effective performance.

c. The administration of an individual test begins as soon as the examinee enters the room. The first step is to get him into the proper mental condition for taking the test; to remove fear and tension which may conceal qualities valuable to the Army. The man reporting for the test is quite apt to be afraid, discouraged, misinformed, or antagonistic and in no condition to perform in a fashion that can be considered a trustworthy indication of his true ability. Here the care exercised in selecting examiners will begin to pay dividends. The skillful examiner will greet the subject in an affable manner, ask him questions about himself, about his work, listen to his complaints, and give every indication of being genuinely interested. He will often have to call upon all his skill and patience to carry this off without creating an air of forced and stilted play acting.

d. The transition from this informal chat to the presentation of the test materials should be gradual and natural. The test manual will contain suggestions for bridging this gap, or such statements as "I have some problems here I would like you to try," or "Let's see what you can do with these questions," will serve the purpose. From this point on, the presentation of the problems, the questions, and all directions to the subject must follow the exact wording of the manual. Moreover, any performance materials, such as blocks or pictures, must be placed on the table or exposed precisely as specified, and the different parts of the test must be given in order, without skipping around. It cannot be overstressed that any departure from the manner of administration in which the test was constructed and standardized will make the test score unreliable.

e. The examiner should speak distinctly and slowly while administering the test so that the examinee may hear and understand, for the examiner may not repeat any questions (unless some unexpected disturbance has prevented his being heard the first time). He must guard against gestures, words, or inflections of the voice that may suggested an answer. Throughout the course of the examination, the examiner will have to motivate the subject to do his best. He will have to make appropriate remarks of approval or praise after each success, and he will have to console and encourage him when he fails. Suggestions for such appropriate remarks will usually be included in the manual for administering the test.

175. Timing

With the individual test, the problem of timing is usually somewhat more difficult than with group tests. This is so because the problems and questions which compose the individual type of test often have separate time limits, and these limits are often specified in terms of seconds rather than large intervals. Furthermore, many of the items of an individual test are scored in terms of the time required to arrive at the correct solution. This means that the examiner will have to look at his watch frequently. Yet, because of the close personal nature of the situation, the examiner cannot be too obviously engrossed in the time problem without creating a disturbing and distracting tension in the examinee. Everyone has experienced the feeling of nervous strain when working against time and the maddening tendency of fingers to become all thumbs as the seconds tick off. Some of this tension is natural, of course, when one is told to work as rapidly as he can; but it is tremendously heightened if the examiner is a nervous clock watcher. So, the timing should be done unobtrusively and with the appearance of casualness. This does not mean, however, that it can be slipshod. It is of utmost importance that the timing be precise and that the exact limits specified in the manual be observed. Where tests are scored in terms of time, an error of a few seconds may account for a difference of several score points. It is essential that the examiner should be thoroughly practiced in timing. If possible he should use a stop watch, because this will always start at zero, and because the timing can be done with one hand, leaving the attention of the examiner focused on the test itself. If a stop watch is unobtainable, an ordinary watch will be used. It should be of the type equipped with a large sweep second hand for ease of reading, and the examiner should always give the starting signal for any item when the second hand is at zero.

Section IV. SUMMARY

176. Importance of Proper Administration

a. The development of an instrument includes the method of administering the instrument. The value of the instrument may be lost if its administration is varied.

b. Proper administration requires careful attention to the following:

(1) Physical surroundings.
(2) Organization and conduct of the testing session.
(3) Answering questions.

c. Administration of individual tests requires more competent administrators than administration of group tests. Because individual testing must seem informal, special precautions are needed to maintain standard conditions.

CHAPTER 12
SCORING ARMY TESTS

Section I. SOME GENERAL CONSIDERATIONS

177. Importance of Following Scoring Instructions

The finest personnel instrument is worthless unless it is administered correctly and unless readings are accurately made from it. Scoring directions are provided in the form of a scoring key; written instructions for its use are usually contained in the instrument manual. Technicians scoring an instrument must understand the procedures and be careful to follow them exactly at all times.

178. Importance of Accuracy

Long experience has shown that, wherever people work with numbers, mistakes can occur very easily. Even the very best workers must be constantly "on their toes" to avoid errors in numbers. Test scores may be used to make recommendations and decisions important to individual soldiers. Sometimes a man's entire military career may be changed because of a high or low score on a test. Therefore, knowing that "most" of the scores on a test are correct is not enough; we must make sure that "every" score is correct. Accurate scoring involves—

a. Complete understanding of scoring keys and scoring procedures.

b. Conscientious following of scoring keys and scoring procedures.

c. Checking of scores obtained.

Section II. SCORING PROCEDURES

179. What is "Scoring"

Basically, "scoring" means comparing the answers which a man makes to the questions on a test with a certain specific pattern of answers, and counting the number of instances of agreement. The number of instances of agreement is called the "score." The specific pattern of answers is called the "scoring key." Usually, the scoring key is merely a way of indicating the correct answer to each question, so that the "score" is then merely the number of correct answers. Exceptions to this are scoring keys for personality and interest tests, where there are no right or wrong answers; the scoring key, in these cases, indicates the answers which are related to the criterion used in validating the instrument.

180. Meaning of the Scoring Formula

a. When the score on a test is the simple count of correct answers (or of the number of instances of agreement with the scoring key), the score may be referred to as a "rights" score. Some tests, for reasons discussed in chapter 2, are scored by counting the number of correct answers and then subtracting from this figure the number of wrong answers, or some fraction of the number of wrong answers. Such scores would be called "rights minus wrongs" scores, or "rights minus one-fourth wrongs" scores, or whatever the case may be. This statement of how the final score is arrived at is called the "scoring formula" for the test. It is stated in the test manual in the section on scoring instructions. It appears in abbreviated form on the scoring key as—

"R" for rights scoring
"R —W" for rights minus wrongs scoring
"R —⅓W"—for rights minus one-third wrongs scoring and so on.

b. Applying the third formula above to the case of the four-alternative multiple-choice items in a test of 100 items, let us see how correction for guessing operates. Let it be assumed that two examinees each know the answers to 50 items of a test, but that whereas one of them stops at this point, the other goes on to make pure guesses on the next 20 items and, by chance, gets five of them right and fifteen wrong. The obtained score of the first examinee will be 50 and that of the second, 55. Application of the scoring formula to both cases, however, will give the first man (50 minus 0) or 50 and the second man (55 minus ⅓ of 15) or also 50. It is important to note that the fraction in the formula depends upon the number of alternatives to each question.* For a test composed of items having five answer choices, the formula would be "rights minus one-fourth wrongs."

c. The scorer must be careful to use the proper scoring formula and to compute it correctly.

Caution: He must be especially careful when certain questions are omitted. The number of wrong answers *cannot* be computed by subtracting number right from total number of test questions; number wrong, in this case, is the sum of number right and number of omits substracted from total number of test questions.

181. Hand Scoring and Machine Scoring

Some tests must be hand scored, and others, when special answer sheets and special pencils are used, may be scored by means of the International Test Scoring Machine. The latter procedure is referred to as "machine scoring." Both methods make use of a scoring formula, but with the scoring machine any deduction for proportion of wrong answers can be made automatically. Not all tests can be machine scored; this is determined by the nature and format of the test. However, all machine-scorable tests may be hand scored. Thus, where for some reason a scoring machine is not available, tests set up for machine scoring can be scored by hand.

182. The Scoring Team

a. Whether scoring is done by hand or by machine, there are a number of steps that must be performed in orderly sequence. Systematic organization of these will produce greater accuracy and reduce scoring time.

b. When a large number of tests are to be scored at one time, efficient scoring will require a team, with a different individual assigned to each of the successive steps in the process. Members of the team will complete only the operations to which they are assigned, passing the papers along to others who perform succeeding operations.

c. With hand scored tests of the type having the answers on the test page itself, it is efficient to have each scorer handle a single page, writing the score at the bottom of the page. Other members of the team should be designated to add together the scores for each page and change total scores into converted scores; still others should check each step in the process.

d. Steps that can conveniently be done by different members of a machine scoring team, of which the operations are described in later paragraphs, are—

(1) Scanning answer sheets.
(2) Operating the scoring machine.
(3) Spot checking scoring by hand.
(4) Hand scoring papers rejected in the scanning process.
(5) Changing raw scores to standard scores.
(6) Checking conversion to standard scores.

e. The importance of checking at all points cannot be over emphasized. If the test is worth enough to be allotted an hour of the examinee's time, it is certainly worth additional seconds to insure that his score is an accurate one.

* The fraction is always $\frac{1}{n-1}$ where n is the number of alternatives to each item.

Section III. HAND SCORING

183. Hand Scoring Test Booklets

Tests in which the answers are made directly on the booklet itself are expendable, as contrasted with nonexpendable tests in which separate answer sheets are used. With tests of the first type, scoring keys, scoring stencils, or a list of correct answers on a strip may be provided. The test manual will contain specific instructions for use of these keys. These instructions must be followed exactly, and all scoring steps checked.

184. Hand Scoring Separate Answer Sheets

Many Army tests make use of separate answer sheets on which all answers are indicated. These test booklets are nonexpendable since they can be reused with new answer sheets. On most of these answer sheets, spaces are provided for possible answers by means of boxes or pairs of dotted lines. The examinee then records his answers by marks in these boxes or pairs of dotted lines. With tests of this type, scoring keys are usually in the form of stencils with holes punched in the positions of the correct answers. Marks showing through the holes when the stencil is placed over an answer sheet are correct answers. Care must be taken to line up the stencil with the edges of the answer sheet, or preferably with two fixed "landmarks" on opposite corners of the sheet. The latter will save time in lining up the key with the answer sheet and will also increase accuracy of scoring. The procedures outlined below for hand scoring separate answer sheets have proved valuable over a period of years and are strongly recommended.

a. Scan (that is, look over) each answer sheet and draw a red pencil mark horizontally through all response positions for each question for which the examinee has indicated more than one choice or no choice. The sum of these red marks will be the number of *omissions.*

b. Place the punch-hole stencil, right side up, over the answer sheet, and line it up with the edges of the paper or the "landmarks." Count the number of marks made by the examinee which appear through the holes of the key, excluding all marks with a red line running through them. The sum of these marks will be the number of *rights.*

c. With the key still in place, count the holes where no mark appears (neither black nor red). The sum of these unmarked positions will be the number of *wrongs.*

d. If the scoring formula is merely rights, only steps *a* and *b* need be performed (step *c* will, however, provide a check, as shown in the following paragraph). If the scoring formula calls for subtraction of a proportion of the number of wrongs from the number of rights, the figures obtained in *b* and *c* will be entered into the proper formula to obtain the raw score.

e. Make the following check: the number of *omissions,* plus the number of rights, plus the number of wrongs should equal the total number of items on the test (exclusive of practice items).

Caution: Sometimes when a standard answer sheet is used, the test will not have as many questions as the answer sheet provides for; in these cases the scorer must be careful to count only the spaces allotted to the test questions.

Section IV. MACHINE SCORING

185. Why Machine Score

a. Past and present widespread use of tests in the Army would be much more difficult without the use of machine scoring. With a test scoring machine, scoring which would require the labor of several men for hours can be done by one man in a much shorter time. Also, machine scoring is generally more accurate than hand scoring.

b. In order to achieve these advantages of speed and accuracy in machine scored tests, three requisites must be met—

(1) Examinees must properly record their responses on the answer sheet.

(2) The scoring machine must be in proper functioning order.

(3) The operator must set up and manipulate the machine correctly.

c. All three requisites can be readily fulfilled —the first by care in administration and proctoring when the test is given; the second, by systematic checking of the machine; the third, by using trained, conscientious operators to do the scoring. The following sections will explain these points further.

186. How the Machine Obtains a Score

The underlying principle of the International Test Scoring Machine is simple. The graphite deposited by a special lead pencil in making a mark on paper will conduct an electric current. If two wires from a source of power are pressed against such a mark, the circuit will be completed. The current will be carried from one wire through the mark to the other wire and it will cause a deflection of the needle of a galvanometer connected in series in the current. If there are hundreds of these simple circuits, all connected to the same galvanometer, all of those which are closed by means of pencil marks will add to the current flowing through the galvanometer. In other words, the amount of the deflection will tell how many of the circuits are completed. In a sense, therefore, the galvanometer reading is a count of the number of pencil marks. If the answers to a test are indicated by pencil marks with graphite in them in a specified place on an answer sheet, and if this answer sheet is then pressed up against a mass of open-end circuits (or electrodes), the dial of the galvanometer will register the number of such marks. But if a punched-hole scoring stencil is inserted between the answer sheet and the electrodes, the current carried by the "right" pencil marks (represented by the punched holes) can be routed one way, and the current carried by the remaining marks routed another way. Thus, the meter dial can be made to register the number of right answers, the number of wrong answers, the number right plus the number wrong, and finally the number right minus any portion of the number wrong —all at the will of the operator and the turn of a switch.

187. IBM Manual of Instruction

It is urgently recommended that all personnel involved in machine scoring of tests become thoroughly familiar with the *Manual of Instruction for the IBM Test Scoring Machine.* Detailed explanations of the principles and operation of the machine, described in the IBM Manual, need not be repeated here. Particular attention is called to the sections of the IBM Manual on "How the Machine Operates," "Description of Machine Parts," "Adjustments for Test Scoring," and "Recommendations for Administering Machine-Scored Tests."

188. Scanning Answer Sheets Before Scoring

a. Once the way in which the machine obtains a score (par. 186) is understood, it is obvious that answer sheets must be marked in certain ways—

(1) Marks must be made on the answer sheet with a special pencil, that is, one with high graphite content.

(2) Marks must be heavy and black enough so that current can flow through them.

(3) There must be no stray marks not intended as answers; any such must be erased since they may also bridge a circuit and produce a score.

(4) Erasures must be clean; remainders of marks may produce a score.

(5) Marks must be placed between the pairs of dotted lines and be as long as the dotted lines. Contacts in the machine line up with these pairs of dotted lines, and marks placed outside the dotted lines cannot complete the electrical circuits.

(6) Only one response should be marked for each question.* Additional marks may be picked up, incorrectly, by the machine as either right or wrong answers, depending on the scoring procedure.

b. Test scorers should not assume that personnel administering the test were able to make sure that every examinee followed these rules perfectly. The first step in the scoring process should be inspecting the papers with the above

* Exceptions:
A few personnel tests call for multiple answers to test questions. Consult specific test manuals for scoring procedures for these tests and keys.

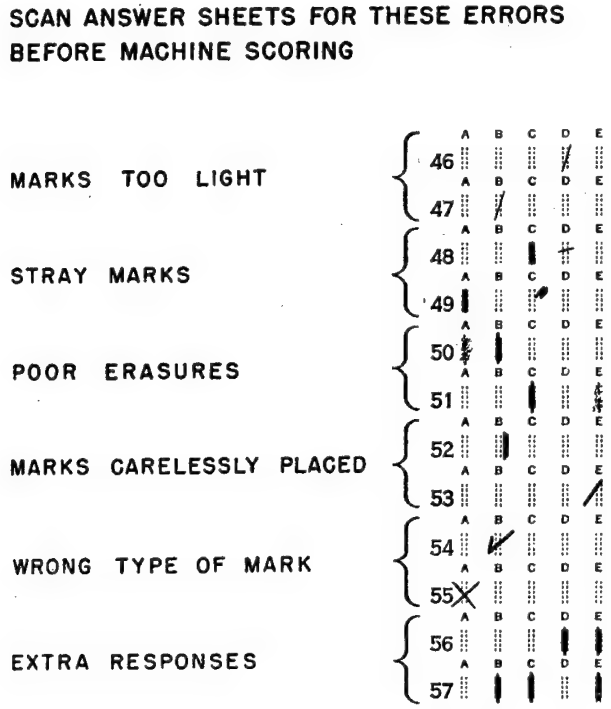


Figure 24. Errors frequently found on answer sheets.

six points in mind; this process of inspection is called "scanning."

c. In most cases, it will be most efficient to lay aside all papers shown by the scanning to be improperly marked; these papers should later be hand scored.

d. Occasionally, a group of papers will fail to meet these requirements in only a few instances; then it may be best simply to "fix up" these few papers, as follows:

(1) Remark, with the proper pencil, responses made with wrong pencils.
(2) Darken marks that are too light.
(3) Erase stray marks.
(4) Clean up poor erasures.
(5) Erase all answers to questions with more than one answer given.

e. From the explanation of the scoring process, it should be clear that only the specially printed answer sheets can be used for machine scoring. The location of the pairs of dotted lines on the paper must be extremely precise so that they will line up with the electrical contacts when the paper is placed in the machine.

189. Setting up and Balancing the Scoring Machine

a. Prepare Check Sheets for the Test to be Scored. Using unmarked answer sheets and a special pencil, mark one answer sheet with all correct responses, mark another so that the score will be approximately the typical score for the group of papers to be scored, mark a third so that the score will be a low score among the group of papers to be scored. For tests where wrongs appear in the scoring formula, the second and third check sheets should contain wrong responses as well as right, and the formula score computed for the check sheets.

b. Place Scoring Stencils in Machine. Remove scoring rack. Lay stencil (or stencils) inside scoring rack as follows: printed side down, wide margin toward open side of rack and slots toward hinges, stencil alined with edges of the rack, holes in stencil alined with holes in rack. Where one stencil only is required (called a rights key), place it inside the top leaf of the scoring rack. Where two stencils are required (rights key and an elimination key), the elimination key is placed inside the top leaf of the scoring rack and the rights key on top of the bottom leaf. Place the scoring rack in the machine, and wind up the key clamp lever.

c. Set Switches on Scoring Machine. Set the master control switch to the field being used (position A, B, or C). Set the formula switch for the field selected to the scoring formula ("R," "W," "R — W," or "R + W").

d. Check the Machine. Insert, in turn, each of the check sheets, prepared as in *a*. In order to score either "R," "W," or "R + W" adjust the "+" rheostat for the field being used so that dial reading corresponds to the score on the first check sheet. Dial readings should then also be correct for other check sheets. In order to score "R—W" "R—⅓ W" "R—¼ W," etc., first see that the separate "R" and "W" scores are read correctly by the machine as just described, then turn the formula switch to "R — W" and adjust the "—" rheostat for the field being used so that the dial reading corresponds to the correct score on the first check sheet.

Dial readings should then be correct for the other check sheets. When the machine is set so that it gives correct scores on all check sheets, leave the formula switch set according to the desired formula. The machine is now ready for scoring.

e. Points To Check When the Machine Does Not Yield Accurate Scores on Check Sheets. The following have been found through experience to be points on which machine operators occasionally "slip up":

(1) Are the scoring keys properly placed in the scoring rack?

(2) Is the scoring rack in proper position in the machine?

(3) Is the key clamp lever wound in?

(4) Are the check sheets marked in the proper positions?

(5) Are marks on the check sheets well made?

(6) Are the check sheets counted correctly?

(7) Are the answer sheets being inserted correctly?

(8) Is the power switch on?

(9) Is the cord plugged to a 110-volt outlet?

(10) Is the master control switch set to the correct field?

(11) Is the formula switch set to the desired formula?

(12) Is the feed channel clear? (No answer sheets accidently stuck half-way through.)

(13) Are the contact blades clean? (These can be cleaned by turning off the power switch and brushing with a long, thin brush.)

Caution: Be sure to turn off the power switch before brushing.

(14) Are the answer sheets completely dry? (A heating unit is built into the machine to dry out answer sheets on humid days.)

Note. If the answer to all of these questions is "yes" and the machine still does not give accurate scores, a serviceman should be called.

190. Scoring the Answer Sheets

When the machine has been checked and set up for scoring the operator inserts answer sheets, one at a time, and records scores in the way specified in the test manual, usually directly on the answer sheet. During the scoring process, the check sheets should be inserted at intervals to be sure that the machine remains in proper balance.

191. Checking Machine Scoring

a. Since the machine can score so rapidly, it is usually possible to have all answer sheets scored a second time by a different operator to check the accuracy of scoring. This is recommended wherever possible.

b. Whether or not a second machine scoring is done, approximately every twenty-fifth answer sheet should be hand scored as a final check.

c. If checking shows agreement in scores, the raw scores are ready to be converted. If scores are found to be in error, all such values should be rechecked by hand scoring. If it appears that a large proportion of the scores are in error, or if there seems to be a consistent error throughout a group of papers, the possibility of machine trouble should be investigated. If this is found and rectified, papers can be rescored by the machine; otherwise all papers must be hand scored.

192. Summary of Steps in Machine Scoring

a. Scan answer sheets.

b. Set up the machine; check the balance of the machine.

c. Score the answer sheets.

d. Check the scoring.

e. Hand score answer sheets which—

(1) Do not "scan" properly.

(2) Do not "check."

Section V. RECORDING SCORES

193. Importance of Accurate Recording

Accuracy and legibility are the most important factors to be stressed in the process of recording scores. Complete, independent checking of numbers copied from one source to another is recommended. Experience has shown, even among careful workers, a certain percentage of number reversals, omissions, repetitions, "switching" names and scores, and other types of errors. Numbers should be typed or hand-written with care. Corrections should be rewritten, *not* traced over the old numbers.

Section VI. SUMMARY

194. Accuracy in Scoring

a. The usefulness of an instrument may be seriously reduced unless proper precautions are taken to minimize errors in scoring.

b. The principal precautions are as follows:

(1) Understanding of scoring keys and procedures.

(2) Careful compliance with instructions on use of keys and procedures.

(3) Systematic checking.

CHAPTER 13

HOW THE ARMY USES PERSONNEL MEASURING INSTRUMENTS

195. Administration and Use of Army Personnel Measuring Instruments

a. The Army administers personnel measuring instruments at various times and places during an individual's Army career. The results may be used immediately to aid in the individual's classification or assignment. They may also be consulted on various occasions during his Army career when such questions as his reassignment or promotion come up. Some typical uses of personnel measurement instruments are shown in figure 25. The table shows where the tests are administered and when and for what purpose the results are used.

b. In addition to instruments listed in figure 25, other personnel tests are available for use. A number of language proficiency tests have been developed to locate individuals possessing critical language skills. A variety of trade tests are available to identify personnel possessing critical occupational skills. A large number of achievement tests have been developed locally and for Army-wide use to measure both individual and group progress in various training programs and schools and as aids in evaluating effectiveness of instruction.

c. Other types of instruments are in use. Ratings in the form of officer efficiency reports are well known. The use of the enlisted efficiency report has been temporarily suspended. An example of the use of a variety of instruments is the selection of enlisted men for Officer Candidate Schools. Ratings, self-description forms, and standardized measurement interviews are used after prior screening by Aptitude Area I tests and the Officer Candidate Test.

196. Classification is a Continuing Process

a. Classification follows a man throughout his Army career, using tests and other measuring instruments. Scores on these instruments and other measures of proficiency are of importance not only to the classification officer at the reception center or the training division but even more to the company commander who must use, to the best advantage, the abilities of the individuals allocated to him. In an emergency, he must have at his fingertips the pattern of the potentialities of each of his men.

b. Classification, however, is not the only element in personnel management in the Army. The classification system is a valuable aid in Army training and the assignment of men to appropriate duties, and it must be integrated with military training and manpower requirements. The realization of the full benefits of an adequate classification system is dependent upon the adequacy of training methods and content and assignment procedures. Manpower requirements and personnel policy decisions affect the usefulness of a classification system. A classification system must, on the other hand, be sufficiently comprehensive and flexible to meet emergencies. Personnel research is one of the principal tools for maintaining the usefulness of the Army classification system.

Tests	Units at which normally administered	UNITS AT WHICH UTILIZED AND FOR WHAT PURPOSE					
		Armed forces examining station	Reception center	Training divisions and replacement training centers	Army schools	Units	Replacement depots
Armed Forces Qualification Test (AFQT).	Armed Forces Examining Stations.	1. To determine acceptability for enlistment or induction as far as mental standards are concerned. 2. To provide a basis for allocating personnel according to mental level to the four services and aid in the equitable distribution of military manpower. 3. To determine whether applicants for special recruitment (OCT, Language, WAC, etc.) meet the mental standards.		A consolidated report by service is submitted to Department of Defense to aid in the proper operation of the program for the qualitative distribution of military manpower.			
Non-Language Qualification Test (NQT).	Armed Forces Examining Stations.	1. To identify those registrants failing AFQT who should be placed in a "special training deferred" category.		A consolidated report is made to Department of Defense and to Selective Service for purposes of planning for full mobilization.			
Verbal-Arithmetic Subtest (AFQT).	Armed Forces Examining Stations.	1. To identify those registrants failing AFQT who should be placed in a "regular training deferred" category.					
Army Classification Battery (ACB) (Test Scores are converted into Aptitude Area Scores).	Reception Conters (may also be administered at civilian components units which have been ordered into active military service).		1. To aid in determination of entry MOS's. 2. To provide a basis for distributing personnel to various training agencies according to mental level; used in conjunction with rate tables. 3. To aid in identifying scientific and professional personnel for assignment to the proper training agency; used with educational qualifications. 4. To determine which personnel should be administered the officer Candidate tests.	1. To aid in school selection and selection for Leaders' course. 2. To aid in selection for special assignment, such as airborne training. 3. To aid in the equitable distribution to units and overseas theaters of personnel who have completed training. 4. To select personnel for common specialist training at Training Divisions. 5. To select personnel for training in various MOS at Replacement Training Centers.	1. To reassign personnel to appropriate courses.	1. To aid in the distribution of personnel to subunits; used with rate tables. 2. To aid in determination of entry MOS. 3. To aid in school selection. 4. To aid in selection of personnel for Leaders' course. 5. To aid in selection of personnel for special assignments. 6. To aid in selection of new potential primary MOS based upon aptitudes.	1. To effect equitable distribution of personnel to units by mental groups within MOS. 2. To aid in selection of new potential primary MOS based upon aptitudes.

Figure 25. Some typical uses of personnel measurement instruments.

Officer Candidate Tests (OCT–1 & 2).	Reception Centers (may also be administered at civilian components units which have been ordered into active military service).			1. To aid in selection for officer Candidate School.		1. To aid in selection for Officer Candidate School.	
-Non-Language Test 2 abc.	EUCOM.	1. To determine whether aliens seeking to enlist in US Army meet minimum mental standards.					
	Armed Forces Examining Stations, Antilles Command.	1. To identify those registrants who should be permanently rejected or placed in a "not presently acceptable" category.					
English Knowledge Evaluation Test (EKE).	EUCOM.	1. To indicate which aliens enlisted in US Army require English language training.					

Figure 25. Some typical uses of personnel measurement instruments—Continued.

APPENDIX I

GLOSSARY OF TERMS WITH INDEX TO RELATED PARAGRAPHS IN THE TEXT

Note. Explanations of terms given here are intended only as aids in reading the text and not as comprehensive definitions of the psychological or statistical concepts involved. For fuller discussion of the terms, see the reference paragraphs listed.

Achievement test—A test of how much a person has learned about a subject or how skilled he has become in a line of work as a result of training and experience. Provides answers to such questions as—Does he do well enough now to be put to work on the job? Has he learned enough in a basic course to go on with advanced training (pars. 19 and 115–126).

Adjectival score—The qualitative statement of the rating of an individual or his performance, such as letter grades (A, B, C, D, E) on school achievement, or descriptions of performance like "excellent," "good," "poor" (par. 67).

Alternate forms—When an instrument is in fairly wide use, it is common practice to have at least two forms, equivalent in content, which have been standardized to yield comparable scores. With such forms, the distributions, means, and standard deviations are almost the same. Often a single conversion table from raw score to standard score is used. Such practice provides a means of retesting men when necessary without their repeating the same test and also helps to prevent the content of a test from becoming familiar to examinees. Alternate forms of an instrument may also be administered to the same group of examinees for the purpose of estimating the reliability of the instrument (par. 82).

Alternative—One of the possible answers supplied to a multiple-choice question of a personnel instrument. The person taking the test indicates in some prescribed fashion which of the alternatives he selects (par. 21).

Aptitude—Readiness in acquiring skill or ability; the potentiality of becoming proficient, given the opportunity and appropriate training. The term may refer to the capacity to learn one specific kind of work or to general trainability (pars. 19 and 106–114).

Aptitude area—Term applied to an occupational area and the personnel measuring instruments associated with Army jobs in that area. The tests in an aptitude area are those which have been found to provide useful indication of chances of success for a number of different jobs. Each set of tests yields a composite score representing a combination of abilities. The jobs in an aptitude area are those for which that combination of abilities has been found to be important. There are at present ten such aptitude areas, each made up of from two to four tests of the Army Classification Battery and the Army jobs for which that set of tests is the best available predictor of success (pars. 109–113).

Aptitude test—A personnel instrument used to obtain an estimate of how well a person can learn to do certain kinds of work or acquire certain skills. This estimate is based on a measure of what the individual's present level is in respect to abilities or skills that have been found to be important in the work for which he is considered. Contrasted with ACHIEVEMENT TEST which measures how well an individual has already learned

to perform a given task. The content of aptitude and achievement tests may be identical; the same test can, and frequently does, serve both purposes. An aptitude test helps to answer such questions as—Can the man be trained in a reasonable length of time to do the job (pars. 19, 108–115).

Arithmetic mean—(*See* Mean.)

Army standard score—A standard score with a mean set at 100 and a standard deviation of 20. Raw scores on Army personnel measuring instruments are usually converted to Army standard scores which state the individual score in relation to the scores of the standard reference population (pars. 73–78).

Associate ratings—Evaluations of a man obtained from those who work with him or who are in fairly close association with him. The term includes superiors, peers, and subordinates. Such ratings have been found to be useful criteria against which to validate various methods of measuring performance (pars. 62 and 158).

Average—A number or value that represents all the values in a series; usually refers to the arithmetic mean of series, but is actually a general term which covers median, mode, and other means (par. 71).

Battery of tests—(*See* Instrument battery.)

Bias—Error in measurement other than chance or random error. In sampling, if each case in the population has the same opportunity to be included, the sample, except for chance errors, will be representative of the population with which a test is to be tried out. However, some influence other than chance may cause a greater proportion of certain types of cases to be included than would result from chance selection; the resulting sample is said to be biased (pars. 48, 61 and 141–147).

Biographical information blank—Same as Self-description form.

Chance—Theoretical probability that an event will occur; variation among samples which causes statistical results obtained on one sample to differ from those on other samples selected on the same basis from the same population (pars. 98–104).

Check-list—A list of items to be noted or verified. In personnel measurement, a list of items, which may serve either as a rating or basis for a rating; they are usually checked to indicate whether or not they apply to the ratee (or his behavior) and sometimes to what degree they apply (par. 154).

Classification—The process by which personnel are evaluated in terms of what military tasks they are fitted to do or to learn to do with a view to their assignment or reassignment to jobs or training. In the Army, classification rests upon an analysis of the individual which takes into consideration his education, experience, interests, and physical status, as well as his aptitudes and abilities as estimated by personnel measuring instruments. Assignment is made in the light of existing manpower needs of the Army. Classification also pertains to the organization of military occupational specialties into related occupational groupings (pars. 12 and 106–114).

Conversion table—A table for changing scores from one kind to another. In the Army, usually a device for interpreting the raw scores earned on a personnel evaluation instrument by translating them into standard score equivalents. Table usually has two columns—a list of all possible raw scores and the standard score corresponding to each raw score (par. 76).

Converted score—Different personnel measuring instruments are likely to yield scores on different scales, with different scale units. A score obtained in terms of one kind of unit which has been translated into the units of another scale of measurement is known as a converted score. In Army practice, results on almost all personnel instruments are converted to Army standard scores (pars. 60 and 159).

Correction for chance success (guessing)—Many persons taking an objective test will answer all the questions whether they know the answers or not. The scoring formula is sometimes set up to deduct from their scores the estimated increase due to guessing. This

is done by subtracting from their score a certain proportion of the number they got wrong (pars. 28 and 180).

Correction for restriction in range—Data available for validating tests in the Army are often obtained not from the entire population of candidates or recruits, but from selected groups of men who have already been picked out for their estimated ability in the very tasks which the tests are assessing. As a result, the groups on which the tests are validated are bound to be closer together in the abilities concerned than the general mass of recruits from which they have been drawn. The result is usually a smaller validity coefficient than would have been obtained with the entire group. Correction gives an estimate of what the validity would be if the entire group were to be tested (par. 30).

Correlation—Relationship or correspondence between two sets of measures. Positive correlation means that persons who stand high in one set also tend to stand high in the other set, and that persons low in one set tend to be low in the other. In an inverse relationship, or negative correlation, persons high in one set of measures tend to be low in the other, and vice versa. Zero correlation means no relationship between two sets of measures (pars. 44 and 88–90).

Correlation coefficient—Numerical expression of the degree of relationship existing between two sets of measures, as, for example, the scores made on two different tests by the same group of people. The coefficient of correlation, abbreviated as "r," cannot be greater than 1 or —1 and is usually expressed as a two-place decimal fraction such as .86 or .08 or —.23. Positive values of r indicate degree of positive correlation; negative values indicate degree of inverse relationship. Values of r cannot be interpreted the same as percentages (pars. 44 and 88–90).

Criterion—The standard of human behavior that a personnel instrument is supposed to measure. In validation, it is the standard against which personnel instruments are correlated to indicate the accuracy with which they predict human performance in some area (pars. 11, 41–62, 89, and 123).

Critical score—(*See* Minimum qualifying score.)

Cross-validation—Repetition of a validation study, using data from a second group of men, for the purpose of seeing whether the validity previously found is maintained (pars. 32, 37, and 140).

Cutting score — (*See* Minimum qualifying score.)

Descriptive Rating—Verbal description of the ratee in the words of the rater. May be limited to designated characteristics or be an over-all estimate of value to an organization, such as the Army. Sometimes used with more objective rating instruments (pars. 121 and 153).

Differential classification—Designation of Army personnel for assignment on the basis of a battery of instruments for evaluating differences in aptitudes and abilities within the individual and between individuals. Implies assignment of personnel on basis of their more promising aptitudes, the final decision being, in part, based upon the military personnel needs of the time and, in part, also upon the education and experience of the individual (pars. 108 and 110–113).

Difficulty (*difficulty index*)—Applied to a personnel instrument or an item of a personnel instrument, difficulty or difficulty index refers to the percentage or proportion of examinees in a representative group who answer an item correctly. The smaller the percentage, the more difficult the item is considered. Difficulty of instruments or items is not a matter of *a priori* judgment but of actual trial, except when used as a rough estimate in selecting items for preliminary try-out (pars. 30, 31, and 118).

Discriminative index—Validity of a test item in terms of its relationship between getting the item right and performance on job or training. If, for example, individuals who perform best on the job tend to answer a given item correctly more often than those who do not perform well, an item is considered valid (par. 30).

Dispersion—The extent to which scores of a distribution spread out from the mean. Dis-

persion measures are in terms of distances on the scale of measurement and show extent of spread. They are used to supplement measures of central tendency, such as the mean, in describing a distribution of scores. The measure of dispersion most commonly used is the standard deviation (pars. 59–61, 71, 72, and 157).

Distortion—In self-description forms, distortion is the result of an individual's tendency, conscious or unconscious, to report his responses so that his score is unduly high or unduly low (par. 143).

Distortion score—A measure of the difference between the score a person should have on a self-description form (as indicated by his standing on the criterion measure) and the score he obtains. Items which are answered in one way by those with high distortion scores and in another by those who give comparatively accurate appraisals of themselves are then used to set up a SUPPRESSOR KEY which allows for such individual variations in accuracy of self-appraisal (par. 143).

Distribution of scores—A tabulation, usually representing a group of scores, showing how many individuals made each score. Scores are arranged in order from highest to lowest. When there are a large number of different score values, the scores are grouped into brackets, or intervals. The frequency for each interval then indicates the number of cases with score values within the interval (pars. 74–76 and 99).

Efficiency rating—A systematic evaluation of performance in an assignment, or of other behavior; usually a rating made by a superior of the ratee (pars. 149–161).

Empirically determined key—A scoring key based on the results of field trial and analysis; distinguished from predetermined key (pars. 90, 123, and 147).

Equivalent forms—(*See* Alternate forms.)

Essay question—A type of question which calls for a complete answer in the form of discussion, comparison, explanation, or identification; an answer which must be thought out and expressed, in contrast to one which is only to be identified, as in multiple-choice questions (par. 121).

Essay rating—(*See* Descriptive rating.)

Essay test—A test composed of one or more essay questions; contrasted with objective test (par. 121).

Expectancy table (chart)—A graph or table, based on a validity coefficient, showing the proportion of examinees earning each score on a predictor instrument or group of predictor instruments who may be expected to succeed in a given assignment. With a high validity coefficient a large proportion of candidates who make high scores will be expected to succeed on the job and a large proportion of candidates who make low scores will be expected to fail. With a low validity coefficient, proportions of probable successes and failures on the job tend to be the same for all scores (par. 89).

External criterion—A criterion that is measured independently of the personnel measuring instrument designed to predict it. Opposed to internal criterion commonly used in computing an internal consistency index (pars. 90, 123, and 147).

Face validity—A personnel measuring instrument which appears reasonably related to job duties and responsibilities is said to be face valid. Term may refer to the opinion that a test or an item is valid, based on the established effectiveness of similar items. A face-valid item or test may or may not prove to have actual validity or relation to job success (pars. 30, 90, 123, and 142).

Follow-up study—Evaluation of progress of men in jobs or training to which they have been assigned on the basis of certain measuring instruments and procedures. The purpose is to determine how well the selection instruments have identified, in advance, the men who would succeed in such jobs (pars. 39, 53, and 62).

Forced-choice item—Usually a type of rating or self-description item in which the subject is required to choose between two or more alternatives which appear equally favorable or unfavorable, but only one of which is

statistically related to the characteristic being measured (pars. 61, 142–147, and 154).

Frequency distribution—(*See* Distribution.)

Graphic rating scale—The scale on which a rating is to be made may be shown as a scaled line (or other diagram showing various steps from high to low) on which the rater records his judgment. Verbal descriptions are usually placed at each scale point to indicate the level or quality of performance which a man should show in order to be placed at that point (pars. 58–61).

Group test—A test which can be administered to more than one examinee at the same time, using the same test directions and duplicate testing materials. Most group tests are paper-and-pencil tests—contrasted to individual test (par. 20).

Halo effect—A tendency in raters to let their general impression about a person influence their judgment concerning specific and independent traits of that person. In operation, halo may result in a person's being rated high on all or most traits or low on all or most traits. It is an important problem in getting valid ratings (pars. 48, 59, and 157).

Individual differences—They way in which persons vary from one another in the pattern of traits into which their behavior is organized. Individual differences is a matter of the varying degree in which different persons manifest traits common to all or almost all persons, rather than of the presence or lack of certain traits. The term may also refer to variations in level from trait to trait in any one person (pars. 3, 108, and 112).

Individual test—A test that can be given to only one person at a time, usually by a single examiner. Administration usually requires a highly trained examiner (par. 20).

Instrument—Any means by which differences among individuals can be measured or the relative standing of the individual in the group determined. Tests, rating forms, inventories, and standard interviews are all personnel measuring instruments (pars. 15–39).

Instrument battery—A group of tests administered for a common purpose. The scores may be used to present a profile for an individual or combined into a single score or rating, with each score weighted according to its contribution to prediction of the criterion, usually success on the job. Generally, a battery can predict performance on the job more accurately than any one of the instruments used alone (pars. 36–39).

Intercorrelations—Correlations existing among two or more sets of measures. Usually refers to correlations among tests or among items (par. 109).

Interest blank or inventory—(*See* Preference blank.)

Internal consistency—Degree to which an item ranks men according to their total score on an instrument. If, for example, the individuals who answer an item correctly tend to answer most other items correctly, while those who answer an item incorrectly receive low total scores, then the item contributes to measuring whatever is measured by the test as a whole, and is said to have high internal consistency. An item answered correctly as often by those who make low total scores as by those who make high total scores does not so contribute and hence is usually taken out of the instrument. The index may take one of several possible statistical forms, but is usually a correlation coefficient (pars. 30–31).

Interview—Conversation between interviewer and interviewee directed along channels predetermined by the purpose of the interview. Purpose may be to give or to get information or directly or indirectly to help in the solution of vocational or emotional problems of the interviewee. In personnel measurement, the purpose of the interview is the evaluation of some aspect or aspects of the interviewee's behavior (*See also* Measurement interview.) (pars. 16 and 127–133).

Inventory—(*See* Self-description form.)

Item (*test item*)—A test question or problem. Any single element of a test to which response is desired. Questions, problems, or tasks comprising an evaluation instrument (pars. 21, 26, 28, 30, 118, and 139–140).

Item analysis—The statistical study of each item of an instrument to find out the extent to which the item contributes to the effectiveness of the test as a whole. It usually includes finding the difficulty, internal consistency, and validity of each item (pars. 30, 140, 143, and 146).

Item analysis key—(*See* Empirically-determined key.)

Item selection—Using item analysis data, the process of selecting from a pool of items those items of specified difficulty and validity to make up the final form of the instrument (pars. 30, 31, 140, 142–144, and 146).

Job—Term used throughout this manual to denote area of work or employment. In the Army often used to mean MILITARY OCCUPATIONAL SPECIALTY, or duty assignment within an MOS (pars. 2, 37, 106–109, and 117).

Job analysis—The process of collecting, evaluating, and presenting detailed information about jobs, including the duties performed, and the knowledges, abilities, skills, and personal qualities required for doing each job satisfactorily. Job analysis is basic to the construction of job proficiency tests. It is helpful in preparing instruments for classification or selection for Army job or assignment (pars. 2, 18, 37, 42, 46, 47, 62, 117, and 118).

Job sample test—An achievement test of performance on an actual job. The job sample test usually consists of a single operation or a sequence of related operations in which the examinee is required to employ the usual materials, tools, and techniques of the job. Sometimes used synonymously with WORK SAMPLE test (par. 116).

Mean—Arithmetic mean. In popular usage is called "average." Computed by adding the value of all scores or measures in a series and dividing the sum by the number of scores or measures in that series (pars. 71–74, 99, 101, and 103).

Measurement interview—An interview conducted according to a uniform procedure so that all the persons interviewed go through, as nearly as possible, the same process. May be conducted by a single interviewer or by a board, members of which observe the interviewee on carefully defined aspects of his behavior and rate him on the basis of their observations. Generally used as one of several instruments in a selection or classification battery. Results in a score which can be used in combination with scores on other instruments (pars. 16 and 127–133).

Military occupational specialty (MOS)—The term used to identify an area of military job activities which require similar skills and abilities for their performance. A military occupational specialty includes duty positions which are sufficiently alike in respect to duties and responsibilities, skills and knowledges, and physical and mental requirements to be adequately described in a single job description (par. 106).

Minimum qualifying score—A score below which candidates are not accepted for assignment. Location of the minimum qualifying score on the scale of measurement depends upon the selection ratio (number of candidates needed for a particular assignment divided by the total number of candidates) and the magnitude of the validity coefficient (pars. 35, 38, 83, and 94–97).

Multiple-choice question—A form of objective test question in which two or more answers, or alternatives, are presented. The examinee is instructed to choose the answer he thinks is right and indicate it in some prescribed way, usually by marking it on a separate answer sheet (par. 21).

Multiple correlation—A technique for determining how to combine tests to get the best possible prediction of a criterion and to estimate what correlation the composite score will have with that criterion (pars. 37 and 89).

Nomination technique—A type of rating in which raters are asked to indicate from among an entire group the persons they consider best and poorest in specified characteristics (par. 60).

Non-language test—A test that requires little or no speaking, reading, or understanding of language on the part of the examinee either

in connection with comprehending directions or making responses. Directions may be given pictorially or in pantomime. Used with illiterates or persons unfamiliar with the language in which tests are given (par. 20).

Non-verbal test—(*See* Non-language test.)

Normal distribution—A distribution of scores or measures such that a normal curve is the best fitting curve. Most measurements of personal traits or characteristics are found to be distributed normally or approximately normally when data is taken for a large and unselected group of individuals (pars. 74, 75, and 79).

Normal probability curve—A frequency curve based on the laws of chance; the form of curve commonly found when a very large number of values are tabulated. Appears as a symmetrical, bell-shaped figure rising above the base-line measurement scale, reaching a maximum height at the center, and tapering to the base-line at both extremes of the scale (pars. 74 and 99).

Normalized standard scores—Standard scores converted so that their resulting distribution is normal. Conversion is usually from raw scores to percentile scores which are then transformed into standard scores (par. 75).

Norms—Norms describe levels of performance reached by specified groups and provide a means of comparing performance of individuals. Norms for Army personnel instruments are usually expressed as Army standard scores and are based on the performance of a standardization sample representing as nearly as possible the population with which the instrument is to be used (pars. 73–78). (*See* Army standard score.)

Objective test—An instrument in which there is little or no disagreement among experts as to the correctness of response and on which the result obtained is almost completely independent of the person doing the testing and scoring, so long as directions are strictly followed (par. 121).

Occupation—(*See* Military occupational specialty.)

P value—Percentage of the group to which a personnel measuring instrument is administered marking a given item or item alternative (par. 146).

Paper-and-pencil test—A personnel measuring instrument, most often verbal, on which the examinee responds to questions by writing or checking his answers, usually on a separate answer sheet. Usually administered to groups, but may be administered to individuals (pars. 20, 116, and 119).

Percentile—The score in a distribution of raw scores below which occur a given percentage of the cases. Hence, the 70th percentile is the raw score below which 70 percent of scores of persons in the group fall (pars. 69 and 74).

Percentile rank—Same as Percentile score.

Percentile score—Indicates the percent of the group which ranks below the specified percentile rank. Thus, an individual who has a percentile rank of 65 exceeds 65 percent of the group and is exceeded by 35 percent of the group (par. 69).

Performance rating—(*See* Efficiency rating.)

Performance test—A test in which the examinee is required to demonstrate some practical application of knowledge or some degree of an essential skill. Frequently, a work sample test, but it may also be in paper-and-pencil test form. More likely to be given individually and to require nonverbal responses. Performance tests are of special value with persons having limited language ability, since the verbal directions may be simple and the response is likely to be a nonlanguage response (pars. 20, 116, and 119).

Personal inventory—(*See* Self-description form.)

Personality questionnaire—(*See* Self-description form.)

Population—All of the cases in the group with which a research study is concerned. In the case of a selection instrument, the population is the total of those available for selection. The term population is sometimes applied, less precisely, to the sample of cases which is taken to be representative of the total

group and on which analysis is carried out (pars. 15, 30, 32, 34, and 98–101).

Power test—Test in which items are usually arranged in order of increasing difficulty and examinees are given all the time they need to complete as many items as they possibly can. Usually contrasted with speed test. Army tests usually measure both power and speed (par. 23).

Practice effect—When a person takes the same test more than once, his scores on the later trials may be higher because of his familiarity with test procedure or content. To reduce this effect, the Army provides alternate forms of measuring instruments for use when retest is necessary (par. 83).

Predetermined key—A scoring key in which answers are scored or weighted on a judgment of how the items should predict performance or behavior, rather than on how they do predict when actually tried out (par. 147). Distinguished from empirically determined key.

Prediction—Estimating criterion performance, such as success on the job, from known performance on another measure or set of measures. The degree of accuracy of prediction can be estimated from the size of the validity coefficient (pars. 44, 45, 55, 62, 83, 85, and 88–90).

Preference blank—A type of self-description instrument designed to appraise systematically the expressed preferences or interests of individuals, usually for specified occupational activities (par. 135).

Probability—(*See* Chance.)

Proficiency test—(*See* Achievement test.)

Profile (*profile chart*)—A graph or diagram which shows a person's relative place in a group on each of several different traits, throwing into relief the pattern of the qualities in which he excels, is average, or in which he is relatively deficient. Such profile patterns may be considered in connection with the known requirements for success on various jobs in deciding upon a suitable assignment for an individual (par. 111).

Random sample—A sample chosen from a population so that every individual case has an equal and independent chance of being included. Does not mean a sample chosen haphazardly without a selection plan. In practice, such pure randomness is seldom approached (par. 100).

Range—The limits within which the actual scores or values on a test or other measurement device fall for a given group. Sometimes expressed as the difference between the lowest and the highest obtained scores, but more often expressed merely in terms of these values, as in the following: "Scores range from 10 to 85" (pars. 31, 59, 61, 65, and 73).

Rank (*rank order*)—In personnel evaluation, the relative standing of an individual, on a given trait, with reference to other members of the group. When all members of a group are arranged in order from lowest to highest, the number 1 may be assigned to the one who stands highest (pars. 60 and 153). (*See also* Percentile score.)

Rating—An evaluation of an individual either as to over-all value or competence or in regard to the degree to which he shows some particular ability or trait; may be in comparison with others in the group or against a fixed standard (pars. 16, 56–63, and 149–161).

Rating form—An instrument on which the rater records his evaluations. Usually combines several rating techniques by means of which an over-all estimation of the ratee's value is reached (pars. 153 and 154).

Rating scale—A device for recording a rater's estimates of an individual, either in respect to specified traits or abilities, or as to his over-all value to an organization. The various levels or degrees which make up the scale may be defined in terms of concrete behavior to be observed. May be accompanied by an actual scale of measurement, as in a graphic rating scale (pars. 56–59 and 157).

Raw score—The score as originally obtained, expressed in the original units of measurement. The total number of right answers, or sometimes the total number of right answers

minus a fraction of the number wrong (pars. 33, 68 and 75).

Regression—In correlation, the tendency of a predicted value to be nearer the average than to the value from which the prediction is made (par. 88).

Reliability—The degree to which an instrument can be relied upon to yield the same result upon repeated administrations to the same individual. Primarily concerned with imperfections within the instrument itself and the method of administering it which result in inaccuracies of measurement. Reliability is usually estimated by correlation between two sets of scores made by the same persons on two equivalent parts of the test or on alternate forms of the same test. It differs from validity in that it is concerned only with the dependability or consistency of the measure and not with whether or not the instrument measures what it is supposed to measure (pars. 11, 22, and 79–83).

Reliability coefficient—An estimate of the consistency of results to be expected from a personnel measuring instrument. Usually a correlation coefficient for two sets of test scores obtained by administering a test twice to the same persons or by administering two equivalent forms of a test to the same group (pars. 82 and 83). (*See* Reliability.)

Sample—Generally refers to a group of individuals taken as representing a larger group; for example, an instrument is standardized on a sample group representative of the operational population with which it will be used. The terms "sample" and "population" are, however, sometimes loosely used interchangeably. "Sample" may also refer to selected items of knowledge or behavior taken as typical of a whole body of knowledge or of many aspects of behavior; for example, a test may "sample" job content (pars. 34, 54, 100, 101, 116, 118, 119, and 123).

Sampling—General process of selecting a limited number of cases from a population (pars. 22 and 98–101).

Scatter diagram—A two-dimensional chart showing the relationship between two measures represented by horizontal and vertical scales. Each point on the diagram represents two measures, and the resulting configuration of points shows the extent and nature of the relationship existing between the two measures (par. 88).

Score (test score)—A general term covering raw scores and converted scores, such as standard scores and percentile ranks (pars. 7, 35, 37, 38, 44–48, 60, and 67–78).

Scoring formula—Part of the scoring directions which states exactly how the final figure expressing the score is to be arrived at. Clarifies such questions as how omissions are to be counted, or whether a portion of the number wrong is to be subtracted to correct for possible success in guessing, etc. (pars. 28 and 180).

Scoring key—The specific pattern of answers to be counted on a measuring instrument to obtain a score (pars. 28, 140, 142, 147, and 179).

Screening—Gross selection early in the total selection process to identify those in the available supply of personnel who meet the minimum qualifications for a given assignment. Selection in which those below the minimum qualifying score on a preliminary evaluating instrument are rejected from further consideration for a particular assignment. Usually takes place early in the selection procedure to avoid subsequent unprofitable testing (pars. 35, 38, 96, and 124).

Selection—The process of choosing a required number of individuals who have the necessary and desirable qualifications for entering upon a certain job or training when the number meeting the minimum qualifications is in excess of the number required. Usually, a broad set of qualities or traits are assessed, and a variety of personnel instruments are employed to do the selecting (pars. 37, 38, 91–95 and 124).

Selection ratio—The number of applicants needed for a given job divided by the total number of applicants available. Along with the validity coefficient, the selection ratio helps to determine a minimum qualifying score on the selection instrument used in selecting personnel for a given assignment.

For example, where the selection ratio is low, that is, when only a small number of applicants are to be selected from many, even a low validity coefficient will help in the selection of candidates who are more likely than not to be successful in the assignment (pars. 91–93).

Self-description form—A technique whereby information is furnished by the individual concerning his background, attitudes, beliefs, and personality reactions. The scoring key must be empirically determined. Score may be useful in predicting his on-the-job success. Usually is in the form of a check-list, inventory, or questionnaire (pars. 16 and 134–148).

Significance (*statistical significance*)—Variations in scores and other measurements may exist either because of chance influences or because of true differences in the trait being measured. Which it is may be inferred by tests of statistical significance from data at hand. Statistical significance refers to the numerical probability that such variations are the result of chance or random errors alone. A variation that is considered "significant" is one which has occurred as a result of factors or circumstances in addition to those of chance. Hence the commonly-heard phrase, "significant at the 1 percent level" means that according to the findings a difference the size of the one obtained, as between two values, might be expected to occur as a result of chance alone one time in a hundred. Statistical significance is a function of the size of a difference, the number of measurements involved, and the spread of scores (standard deviation) of the distribution. A difference may be small but still be statistically significant because of the large number of measurements involved. On the other hand, variation among measures may be apparently large but still not significant statistically because the number of measurements involved is low. How much consideration should be given to statistical significance depends upon the conditions under which a measure is to be used (pars. 98–104).

Speed test—An instrument measures speed of performance if the time limit is set so that almost no one can finish all the items or tasks making up the test (par. 23). Opposite of Power test.

Spread—(*See* Dispersion.)

Standard deviation—A measure of the spread of scores or how much members of the total group vary among themselves. Provides a means of knowing how far above or below the group average each individual measure falls, in relation to the other measures. The more the scores made by members of a group tend to bunch together, the smaller will be the standard deviation. In a normal distribution, the distances along the scale of measurement laid off by standard deviation units on either side of the mean can be expressed in terms of proportions of the total number of cases. Computed by squaring the deviations from the mean, averaging them, and then taking the square root of the result (pars. 72, 74, and 99).

Standard score—A score expressed in terms of the number of standard deviation units by which it differs from the group average. It enables any individual's score to be compared to the performance of the whole group and thus be given meaning (pars. 70–77). (*See* Army standard score.)

Standardization—The administration of a test or other personnel evaluation instrument to a sample representative of the population with which it is to be used to determine the proportion of the group that will reach or exceed each test score. It is then possible to compare scores made by individuals on subsequent administrations of the test with the performance of the standardization population (pars. 33–35, 69–77, and 159).

Stress interview—An interview in which the interviewee is subjected to specifically-planned emotional stress, the purpose being to permit observation of his characteristic modes of response to such stress, such as his degree of emotional control. Techniques of the stress interview are not appropriate in either the information or counselling type interview, nor are they commonly employed in the usual measurement interview (par. 16).

Subject-matter specialist (subject-matter expert)—An individual expert in some occupational or job area in which tests are constructed. He may act in an advisory capacity regarding achievement test content or may be trained to construct the actual test items (pars. 26 and 123).

Suppressor key—A key used in scoring self-description tests to adjust an individual's score for any constant tendency to select responses according to face validity (pars. 142 and 143).

Test battery—(*See* Instrument battery.)

Test item—(*See* Item.)

Test plan—An organized outline of the knowledges, skills, abilities and attitudes needed to perform a particular job successfully. Each part is weighted in proportion to its importance to the job. It is from such a content outline that an achievement test should be constructed (par. 118).

Test score—(*See* Score.)

Validation—The process of trying out a personnel measuring instrument to determine its effectiveness in predicting performance on an assignment. This effectiveness is expressed in terms of the correlation between scores on the instrument and scores on a criterion of proficiency in job or training (pars. 29–32, 42, 44, 123, 140, 147, and 158). (*See* Validity and validity coefficient.)

Validity—In general, the extent to which a measuring instrument really measures the skill, area of knowledge, aptitude or other characteristic it is supposed to measure. The extent to which a personnel evaluation instrument is able to select in advance persons who will do well in actual assignments, and detect those who are likely to do poorly. A test valid for one purpose may not be valid for another purpose. Validity, hence, is not an intrinsic quality of a measuring instrument but is relative to the purpose of the instrument (pars. 11, 30, 31, 42, 88–90, 123, and 141–146).

Validity coefficient—A statistic that shows the degree of relationship between scores on an instrument and performance in a particular job or training program. It is usually a coefficient of correlation between a test and a criterion the test is intended to predict. Indicates the extent to which individuals who score high on the test also score high on the criterion and those who score low on the test score low on the criterion (par. 89).

Variability—(*See* Dispersion.)

Variable—Any measure which can take on different or graduated numerical values, such as age or scores on an evaluation instrument (par. 88).

Verbal test—Theoretically any test in which language is involved. In general usage the term is restricted to those tests in which the questions and responses are mainly expressed in language or which use language to a substantial degree (par. 20).

Weighting—The process of determining, either by judgment or by statistical means, the relative importance each test in a battery or each item in a test should carry in the overall or composite score (pars. 37 and 47).

Work sample—A small problem representative of the job as a whole, chosen and adapted for the purpose of testing performance on important operations of the job as nearly under normal conditions as possible apart from an actual try-out on the job. Performance on a work sample is frequently used as a criterion against which prediction devices in personnel evaluation are validated (pars. 20, 54, and 116).

APPENDIX II

SELECTED REFERENCES

The following selected references are provided for those who wish to increase their knowledge of the technical principles and methods described in this manual. Other books, of course, may be used.

PERSONNEL MEASUREMENT METHODS

1. Adkins, D. C., *et al., Construction and Analysis of Achievement Tests.* U. S. Government Printing Office, Washington, D. C., 1947. $1.50.
2. Goheen, H. W., and Kavruck, S., *Selected References on Test Construction, Mental Test Theory and Statistics,* 1929-1949. U. S. Government Printing Office, Washington, D. C., 1950. $1.50.
3. Thorndike, R. L., *Personnel Selection.* John Wiley and Sons, Inc., New York, N. Y., 1949. $4.00.

GENERAL PSYCHOLOGICAL STATISTICS

1. Garrett, H. E., *Statistics in Psychology and Education.* Longmans, Green and Co., New York, N. Y., 1947. $4.00.
2. McNemar, Q., *Psychological Statistics.* John Wiley and Sons, Inc., New York, N. Y., 1950. $4.50.
3. Walker, H. M., *Elementary Statistical Methods.* Henry Holt and Co., New York, N. Y., 1943. $3.75.
4. Walker, H. M., *Mathematics Essential for Elementary Statistics.* Revised edition. Henry Holt and Co., New York, N. Y., 1951. $3.00.

☆ U. S. GOVERNMENT PRINTING OFFICE: 1953—200628